The American Economic Review

ARTICLES

MARCH 1986

THE AMERICAN ECONOMIC ASSOCIATION

• Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

• *THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1986, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- \$37.50 if annual income is \$30,000 or less;
- \$45.00 if annual income is above \$30,000, but no more than \$40,000;
- \$52.50 if annual income is above \$40,000;
- \$18.75 annually for registered students (three years only).

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), \$105 a year; individuals, \$70.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$12.00 to the annual rates above to cover extra postage.

• Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

• Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

• Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

// THE AMERICAN ECONOMIC REVIEW

• Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 169 Nassau Street, Princeton, NJ 08542-7067.

• Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

March 1986

VOLUME 76, NUMBER 1

Articles

THE AMERICAN ECONOMIC REVIEW

• Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:
Orley Ashenfelter, Managing Editor, *AER*; 169 Nassau Street, Princeton, NJ 08542-7067.

• Submission fee: $25 for members; $50 for nonmembers. Style guides will be provided upon request.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

March 1986

VOLUME 76, NUMBER 1

Articles

Shorter Papers

Charles P. Kindleberger

International Public Goods without International Government†

By CHARLES P. KINDLEBERGER*

When the word of my prospective elevation to this exalted position first circulated at MIT at the end of March 1983, I happened to encounter Peter Temin in the library. He offered congratulations, and added: "In your presidential address, skip the methodology. Tell them a story." This is the technique that he and Paul David used to great effect in the session on economic history at Dallas a year ago. I choose, however, to follow the lead of another economic historian, Donald McCloskey, who maintains that economics should be a conversation (1983).

In a recent paper, unpublished I believe, George Stigler discussed "the imperialism of economics," which, he claims, is invading and colonizing political science—through public choice theory and the economic theory of democracy—law, and perhaps especially sociology, where our soon-to-be president-elect, Gary Becker (1981), has extended the reach of economics into questions of the family, marriage, procreation, crime, and other subjects usually dealt with by the sociologist. "Imperialism" suggests super- and subordination, with economics on top, and raises the question whether as a profession we are not flirting with vainglory.

My interest has long been in trade, and I observe that economics imports from, as well as exports to, its sister social sciences. In public choice, we can perhaps explain after the event whose interest was served by a particular decision, but we need political science to be able to forecast which interest is likely to be served, whether that of the executive, the legislature, the bureaucracy, some pressure group—and which pressure group — or, in the odd instance, the voters. Individuals act in their own interest, let us grant, but a more general motive of emulation may be drawn from sociology as Adam Smith was aware in the *Wealth of Nations* (1776, p. 717), as well as in *The Theory of Moral Sentiments* (1759 (1808), I, p. 113). I want today to borrow one or two ideas from political philosophy, and to conduct a conversation with a new, impressive, and growing breed of political scientists working on international economic questions. The discussion falls into two loosely connected halves—the first dealing with what economists can, perhaps should, and to some extent do, import from political philosophy and sociology; the second dealing more especially with international public goods.

I

That sharp and sometimes angry theorist, Frank Graham (1948), thought it a mistake to think of trade between nations. Trade took place between firms, he insisted. The fact that they were in different states was irrelevant so long as economic policy was appropriately minimal, consisting perhaps of free trade, annually balanced budgets, and the gold standard. But states may differentiate between firms, through such measures as tariffs, embargos, monetary, fiscal, and exchange rate policy which affect all firms within a given space, and this adds a political dimension (see my 1978 study). The essence may go deeper. In an early graduate quiz, I asked for the difference between domestic and international trade, expecting a Ricardian answer on factor mobility. One paper, however, held that domestic trade was among "us," whereas international trade was between "us" and "them." The student who

*Ford International Professor of Economics, Emeritus, Massachusetts Institute of Technology and Visiting Sachar Professor of Economics, Brandeis University; Box 306, Lincoln Center, MA 01773. I have benefited from comments and suggestions on an earlier draft from Susan Okin, Walt W. Rostow, Walter S. Salant, and Robert M. Solow. A paper with the same title before translation, but with a different coverage was written in 1980, and has appeared in French (1985).

†Presidential address delivered at the ninety-eighth meeting of the American Economic Association, December 29, 1985, New York, NY.

wrote this (now escaped from economics and teaching international law at a leading university) had come from Cambridge University and a course with Harry Johnson. We go beyond this simple statement today in saying that nations are groups of people with common tastes in public goods (Richard Cooper, 1977). Geography discriminates between countries, as a hypothetical customs union between Iceland and New Zealand would demonstrate, and so do governments. Behind and alongside of governments, people discriminate.

Public goods, let me remind you, are that class of goods like public works where exclusion of consumers may be impossible, but in any event consumption of the good by one consuming unit—short of some level approaching congestion—does not exhaust its availability for others. They are typically underproduced—not, I believe, for the Galbraithian reason that private goods are advertized and public goods are not—but because the consumer who has access to the good anyhow has little reason to vote the taxes, or pay his or her appropriate share. Unless the consumer is a highly moral person, following the Kantian Categorical Imperative of acting in ways which can be generalized, he or she is apt to be a "free rider." The tendency for public goods to be underproduced is serious enough within a nation bound by some sort of social contract, and directed in public matters by a government with the power to impose and collect taxes. It is, I propose to argue in due course, a more serious problem in international political and economic relations in the absence of international government.

Adam Smith's list of public goods was limited to national defense, law and order, and public works that it would not pay individuals to produce for themselves. Most economists are prepared now to extend the list to include stabilization, regulation, and income redistribution (Cooper, 1977), even nationalism (Albert Breton, 1964), and standards that reduce transaction costs, including weights and measures, language, and money. Public goods were popular a decade ago. There is something of a tendency today, at least in political science, to draw back and claim that such institutions as open world markets are not public goods because countries can be excluded from them by discrimination. One monetarist goes so far as to maintain that money is not a public good, arguing, I believe, from the store-of-value function where possession by one individual denies possession by others, rather than from the unit-of-account function in which exclusion is impossible and exhaustion does not hold (Roland Vaubel, 1984).

II

Before addressing international public goods, I want to digress to suggest that there are other limits to the imperialist claims of economics. Social goods are not traded in markets, for example—honor, respect, dignity, love. In his address to the Columbia University Bicentennial Assembly, Sir Dennis Robertson asserted that what economists economize is love (1955, pp. 5–6). Michael Walzer (1983, pp. 101–02) has compiled a list of "things" that contemporary moral philosophy will not tolerate being bought and sold: human beings, political power, criminal justice, freedom of expression, marriage and procreation rights (*pace* Becker), the right to leave the political community, exemptions from military service and jury duty, political office, basic services like police protection, desperate exchanges such as permission for women and children to work fourteen hours a day, prizes and honors, love and friendship, criminally noxious substances such as heroin. The inclusion of a number of items on the list is debatable, and history reveals that most of them have been traded on occasion in some cultures. The market, moreover, strikes two lawyers as a dubious device for making "tragic choices," like those in which scarcity confronts humanistic moral values, for example, allocating food in famine, children available for adoption, or organ transplants (Guido Calabrese and Philip Bobbit, 1978). It is difficult to dissent from Walzers's conclusion that a radically *laissez-faire* economy would be like a totalitarian state, treating every social good as if it were a commodity (1983, p. 119). There is, moreover, a similar remark from a

founder of the Chicago school, Frank Knight, who said that the extreme economic man, maximizing every material interest, and the extreme Christian, loving his neighbor as himself, were alike in that neither had any friends.[1]

To admit social goods, not traded in markets, into our economic calculus does not call for altruism. Economists are reluctant to depend on self-denial to any degree (Kenneth Arrow, 1975, p. 22), and moral philosophers are not far behind. To a modern student of ethics, James Fishkin (1982, ch. *ii*), obligations to others fall into three categories: minimal altruism, where the benefit to the receiver is substantial and the cost to the altruist low—the acts of a cheap Samaritan; acts of heroic sacrifice that are not called for; and a robust zone of indifference where one has no cause to be concerned over the effects of one's acts on others. This is for positive actions. Acts that harm others are proscribed by the Golden Rule. Adam Smith expressed the same viewpoint forcefully: "Every man is, no doubt, by nature first and principally recommended to his own care" (1759 (1808), I, p. 193), but goes on: "Although the ruin of our neighbour may affect us less than a very small misfortune of our own, we must not ruin him to prevent that small misfortune, or even to prevent our own ruin" (ibid., p. 194). Does this prohibit us from playing zero-sum games or negative non-zero-sum games? In international trade, must we refrain from levying the optimum tariff? The optimum tariff works to self-interest mainly in the absence of retaliation, and if Adam Smith excludes hurting our neighbor, he recognizes that "as every man doth, so shall it be done to him, and retaliation seems to be the great law of nature" (ibid., p. 191).

Note parenthetically that today's moral philosophers cover a wide territory either side of Fishkin, from Peter Singer (1972) at one extreme whose criterion of justice requires successive acts of altruism until the welfare of the recipient has risen to that of the giver which has fallen, to Robert Nozick (1974) at the other who believes that self-interest rules out altruism almost altogether.

III

Self-interest then is legitimate over a large zone of indifference provided that justice is served by our not hurting others. But the robust zone of indifference applies to strangers, and not to those with whom we have a special relationship, sharing collective goods. It does not apply in the family, the neighborhood, in clubs, in the tribe, racial or religious group, or in the nation. There is some uncertainty whether it applies in regions within a country—New England, the West, the South—or to arrangements between countries short of the world level such as North America or the European Common Market. Collective goods involved here are distributed by mechanisms different from the market: gifts, grants, unequal exchange, sharing through a budget according to need, interest-free loans, inheritance, dowries, alimony, and the like all have a place. Membership in these groups is decided in various ways: by birth, by choice—as in moving into a certain neighborhood or migrating between countries, by application for admission and acceptance. Walzer defends the right of countries to keep out would-be immigrants motivated by economic self-interest, but not those subjected to persecution: "The primary good that we distribute to one another is membership in some community" (1981; 1983, ch. *ii*, p. 1). He argues, however, that states lack the right to keep members from emigrating if there is some other community ready to take them in. Clubs discriminate against outsiders. Neighborhoods are more complex, being presumably open to anyone able to afford and find a place to live, but, in sociological reality, often exhibiting tendencies to attract their own kind and repel others, including harassment or unwritten or even legal restrictions against property ownership. The groupings are amorphous, but they exist.

The nature of the positive bonds that link families, neighborhoods, tribes, regions, and

[1] This at least is oral tradition. I have been unable to find a specific reference in Knight (1936), or Knight and Thornton Merriam (1947).

nations is usually taken for granted and left unexplored, but the consequences are not. Albert Hirschman (1970), for example, makes a distinction between voice and exit: voice —speaking up and trying to persuade—being the appropriate action when one disagrees with the course followed by a group to which one belongs; and exit—resigning or refusal to buy the good or service—as a response to what one dislikes in the market. Adam Smith minimizes the difference between families and strangers, suggesting that affection is little more than habitual sympathy produced by propinquity; despite the greater thickness of blood than water, he claims that siblings educated at distances from one another experience a diminution of affection (1759 (1808) II, pp. 68–70). In arguing against Walzer's view that countries owe immigrants the right to become citizens, Judith Lichtenberg (1981) echoes Smith's view in saying that the crucial difference between members and strangers lies between those with whom one has face-to-face contact and those with whom one does not. An accident that kills someone in one's town or a neighboring community is likely to be more moving than a catastrophe at the other end of the world in which hundreds or thousands die. Adam Smith goes further, comparing the loss of a little finger with a catastrophe that swallowed up China: "...if he lost his little finger he could not sleep, but for China he can snore...provided he has never seen them" (ibid., I, p. 317).

Some years ago in a book on the brain drain, Harry Johnson (1968) argued in favor of a cosmopolitan solution, encouraging emigration, and Don Patinkin (1968) for a national one. In discussing the Bhagwati scheme for taxing professional emigrants earning more abroad than at home, for the benefit of the poor sending country—saying this was akin to paying alimony in a divorce case for breaking a social taboo—I suggested (1977) that the Johnson position was equivalent to saying that a person should go where he or she could earn the highest return, while Patinkin said that people should stay where they belonged. Patinkin chided me privately for this interpretation, and it is admittedly oversimplified. But the difference between the Johnson and the Patinkin positions, both emanating from Chicago, suggests the line between market and nonmarket areas in economics is shadowy.

In writing about the multinational corporation, I have from time to time suggested that host countries resist the intrusion of strangers because "...man in his elemental state is a peasant with a possessive love of his own turf; a mercantilist who favors exports over imports; a Populist who distrusts banks, especially foreign banks; a monopolist who abhors competition; a xenophobe who feels threatened by strangers and foreigners" (1984, p. 39), usually adding that it is the task of international economics to extirpate these primitive instincts and to teach cosmopolitanism. The fact that some of these reactions remain at a late stage in the educational process can be tested by the device of asking students on examinations, *seriatim*, a series of questions:

Do you advocate free trade, or at least is there a strong presumption in its favor?

Do you advocate the free international movement of portfolio capital?

...of corporate capital in foreign direct investment?

...free migration of students and professional labor?

...immigration of relatives of persons permanently resident in this country?

...free migration for all?

(It is desirable to feed these questions to the victims one at a time, without revealing the whole list before the first answer is given, and to take up the replies to the first questions so that there is no chance to go back and amend early answers.) There will be sophisticated answers expatiating on the second, third, and fourth-best if the marginal conditions for a Pareto optimal solution are not met, and I would particularly excuse a James Meade (1955) solution that would limit immigration from countries that have not accomplished their Malthusian revolution, on the ground that their emigrants will be replaced, so that free immigration will reduce world income per capita, if not world income as a whole. Most economists and noneconomists alike would agree, however, that goods are less intrusive than money, money

less so than corporations with control over *our* economic decisions.[2] Intellectuals with whom we identify are hardly intrusive at all. Most of us grant that relatives must be permitted to come together. On the other hand, free migration of labor in general poses a threat to the national identity. The Swiss cut off immigration, despite the appeals of business for more labor, when immigrants constituted one-third of the labor force. In Germany, separate localities felt threatened and stopped inward migration when immigrants reached 12 percent of the resident population. Feelings differed, of course, depending upon the origin of the migrants and their appearance, language, and religion.

One early venture of international economics into this line of investigation was Robert Mundell's "optimum currency area" (1961), initiating a discussion of how large the area for a single currency should be, that can readily be extended to economics in general and to other social sciences. Mundell defined an optimum currency area as one where labor moved freely within the area, but not between it and other areas, taking us back to the Ricardian criterion distinguishing domestic from foreign trade: factor mobility within but not between countries. In neither case is the discontinuity in mobility explained. Perhaps something is owed to low transport costs, but additionally, factor mobility requires a group with such strong social cohesion that those moving are willing to shift, and those at the receiving end are content to receive them.

Ronald McKinnon (1963) offered a different criterion: an optimum currency area was one that traded intensively at home, but only to a limited extent abroad. This implied that tastes within a country are homogeneous for traded goods (as well as for public goods), and that regionally specialized production had grown up to serve those tastes. The Mundell and McKinnon criteria do not necessarily converge: on Mundell's standard, Canada is too big to be an optimum currency area, because of limited movement between Quebec and the English-speaking parts of Canada, and the comparative isolation of the Maritimes and Vancouver. On McKinnon's criterion, however, it was too small because so much of its trade is with the United States.

If one broadens the issue from the optimum currency area to economics more generally and to the other social sciences, anomalies arise from the divergence between the optimum economic area, which on efficiency grounds I take to be the world, and the optimum social unit, one that gives the individual a sense of belonging and counting —which is much smaller. In shifting to the optimum political unit, at least two problems arise, one related to the nature of the ties, the other to the ambitions of its members. To take the second point first, for a nation bent on glory—led by a Bismarck or a de Gaulle —bigger is better; whereas if one is merely trying to get along without trouble, like, say, Denmark, small is beautiful enough.

On the first issue, political ties vary widely. There are leagues, alliances, commonwealths, confederations, federations, provinces, states, principalities, kingdoms. Some lesser units are "united" in varying degrees, as in the United Provinces of the Netherlands, the United States of America, the United Kingdom of Great Britain, and Northern Ireland. The North in the American Civil War was a union, as the Union of Socialist Soviet Republics asserts it is. The small amount of literature I have explored in examining the differences among these forms is not very conclusive, but perhaps the main distinction is between a single state that is centralized, and federations that are loosely joined, with greater powers at the local level. Designations are not always congruent with reality: the Federal German Republic is highly unified, despite the efforts of the occupation powers after World War II to spread politi-

[2]If the intrusiveness of goods is less than that of corporations from abroad, it is perhaps anomalous that the standard of friendly international dealings exemplified in treaties of Friendship, Commerce, and Navigation is less hospitable for goods than for corporations. Foreign corporations in theory are given national treatment; goods only that of the most-favored nation. In practice, many countries ignore the commitment to national treatment and discriminate both against foreign corporations as a class, and among those of different nationality.

cal power widely; the Federal Reserve System was created as a loose agglomeration of twelve regional money markets but quickly fused into a single system in World War I. Centralization and federalization have reflections in demography and in finance. City populations in unified states follow a Pareto-skewed distribution with a single dominant city like London, Paris, or Vienna, and no close rival among the tail of smaller cities and towns. In federations the distribution of cities is log normal (Brian Berry, 1961). Parallel to the demographic division is the financial. Paris has 91.3 percent of French bank clearings; London 87 percent of those for Britain. The contrast is with Canada: Toronto, 37.3 percent; Montreal, 25.5 percent; Vancouver, 6.5 percent. Between these extremes lies Japan with Tokyo 51.2 percent and Osaka 19.7 percent (Jean Labasse, 1974, pp. 144–45).

One explanation for differences between centralized and federal states is historical: where larger states were formed later from unification of lesser units, administrative and financial functions were already being discharged at the local level, reducing the need for centralized services. This hypothesis faces the difficult counterexamples of Italy and Germany, unified out of smaller units in the second half of the nineteenth century, that quickly centralized administrative and financial functions, in Rome and Milan for Italy, and in Berlin for Germany. Another explanation runs in terms of size, with larger states necessarily federal because of the difficulty of providing administration to local units over long distances. This fits Canada, Australia, the United States, perhaps India, but fails to account for Switzerland, unless size is a proxy for maintaining a dense network of communication, and division of valleys by high mountains produces barriers equivalent to those of continental states. If the mathematically minded among you need an analogue, think of federal states as decomposable matrices.

The difference between a single state and a federation may be illustrated with two examples. Some years ago, Seymour Harris (1952) wrote a book on New England in which he claimed that the area got a raw deal from the rest of the country because it paid more in taxes to the federal government than it received in federal expenditure. This thesis implicitly violated the distinction between a budget and a market: in a market equal values are exchanged. A budget, on the other hand, is a device expressing the cohesion of a sharing group with monies raised according to one standard, perhaps ability to pay, and expenditure distributed according to another, some combination of efficiency and need. The other example, equally shocking to an international trade economist, was the notion of the *juste retour*, or fair return, propounded by France in connection with expenditure for joint projects in Europe. France insisted that all monies contributed by her be spent in France. Tied sales are a third- or fourth-best device to limit balance-of-payments deficits for a given contribution to joint efforts, or to maximize the contribution for a given deficit. They are inefficient rather than fair.

IV

But I want to move on to the geopolitical unit that produces public goods. It is a cliche that these have increased in size as costs of transport and communication have declined. Under the eighteenth-century Poor Law in England, the parish resisted immigration from neighboring parishes because of reluctance to share with outsiders. Fernand Braudel (1982) and Sir John Hicks (1969) have each expatiated on the rise of the size of the economic unit from the city-state to the nation-state. National and international markets for goods and money grew slowly, with entrepot centers that intermediated between buyers and sellers surviving in money—cheap to move in space—and largely disappearing for goods where costs of transport were high and could be saved by direct selling, rather than relaying goods through fairs in the Middle Ages and later through cities such as Amsterdam, Hamburg, Frankfurt, and London. The hub-and-spoke system recently discovered in airplane travel and still in place for money has long been superceded in goods. Caroline Isard and Walter Isard's (1945) point that the most pervasive changes in the

economy came from innovations in transport and communications remains valid: contemplate the rudder (in place of the steering oar), fore-and-aft sails; the turnpike; canal; railroad (despite Robert Fogel, 1964); the steamship; iron-clad ship; telegraph; telephone; refrigerator ship; radio; airplane; bulk carrier; jet airplane; satellite television. The numbers of people brought into face-to-face contact across continents and hemispheres has increased exponentially. It is true, to be sure, as was said about a well-known governor and presidential candidate, that it was impossible to dislike him until one got to know him, and increases in mobility and communications have been accompanied by separatism: of the Walloons from the Flemish in Belgium, of Scotland and Wales in the United Kingdom (to pass over the troubled Irish question), and of the *Québecois* in Canada.[3] But it is easier than in Adam Smith's day to imagine ourselves in the circumstances of the Chinese, the inhabitants of the Sahelian desert in Africa, or the tornado-struck islands of Bangladesh as we see them nightly on our television screens via satellite. Do wider communication and transport change the production and distribution of public goods?

Conflicts between economics and political science abound, and many arise from the fact that goods, money, corporations, and people are mobile, whereas the state is fixed. The increase in mobility produced by innovations in transport and communication during and after World War II led some of us to conclude that the nation-state was in difficulty. A reaction occurred in the 1970's. It is significant that Raymond Vernon's influential book *Sovereignty at Bay* (1971), showing the multinational corporation ascendant over the state, was followed by his *Storm over Multinationals* (1977) in which the position is reversed. Cooper's *The Economics of Interdependence* (1968) was followed by an upsurge of interest in national autonomy, decoupling, and pluralism among political scientists, most of whom approve the nation-state and have as heroes, if they will forgive me, not Adam Smith and Woodrow Wilson, but Otto von Bismarck and perhaps even Charles de Gaulle. The tension remains, however. Mobility limits the state's capacity to enforce its writ in taxation, in foreign policy, in standards on such matters as antitrust, pure food and drugs, insider trading in securities, and the like. Mobility undermines social cohesion through the easy intrusion of different nationalities, races, religions, and traditions into the body politic.

V

I come at long last to international public goods. The primary one is peace. Economists are poorly qualified to discuss how, after war, peace is restored and maintained. Most of us reject the Marxian view that war grows directly out of capitalism, and as ordinary citizens and amateur students of history are prepared to agree that peace may be provided by a dominant world power—Pax Romana or Pax Britannica—or by balance-of-power maneuvering, although that seems accident prone. Among the more audacious economists producing an economic theory or set of theories on war is Walt Rostow (1960, pp. 108 ff.). There are views that ascribe war to population pressure, to ambitious rulers aggressively seeking power, and to complex miscalculation. How these are to be avoided or contained is a question primarily for political science.

In the economic sphere, various international public goods have been identified: an open trading system, including freedom of the seas, well-defined property rights, standards of weights and measures that may include international money, or fixed ex-

[3] Tastes in public goods can of course differ within countries. A striking comparison is furnished in E. Digby Baltzell's *Puritan Boston and Quaker Philadelphia* (1979). Boston is characterized as intolerant, extremely homogenous, ascetic, philanthropic, and devoted to social and political responsibility. Philadelphia, on the other hand, was an ethnic and religious melting pot, materialistic, believing in money making, and shunning power and responsibility. Boston produced four presidents of the United States, including one non-Puritan affected by the values of the city, Philadelphia none. Social scientists are wary of ascribing social responses to national (or urban) character. There may nonetheless be occasions when it is inescapable.

change rates, and the like. Those that have interested me especially in a study of the 1929 depression and other financial and economic crises have been trading systems, international money, capital flows, consistent macroeconomic policies in periods of tranquility, and a source of crisis management when needed. By the last I mean the maintenance of open markets in glut and a source of supplies in acute shortage, plus a lender of last resort in acute financial crisis (see my 1973 book, revised 1986, forthcoming).

Public goods are produced domestically by government, unless the governmental agenda is blocked in stalemate among competing distributional coalitions as described by Mancur Olson (1982). Voluntary provision of public goods is plagued by the free rider. In the international sphere where there is no world government, the question remains how public goods are produced. Ralph Bryant is one of the few economists who has discussed the public good element in international cooperation. His vocabulary is different from that of the political scientists: their "regimes" are his "supranational traffic regulations" (1980, p. 470), and he expects leadership in cooperation in monetary and fiscal policy from supranational institutions such as the International Monetary Fund (p. 481). I find this doubtful on the basis of the interwar record of such institutions as the League of Nations.

Political science in this field has produced two schools: the realists who hold to a national-interest theory of international politics, and the moralists, whom Robert Keohane prefers to call "institutionalists" (1984, p. 7). Realists maintain that international public goods are produced, if at all, by the leading power, a so-called "hegemon," that is willing to bear an undue part of the short-run costs of these goods, either because it regards itself as gaining in the long run, because it is paid in a different coin such as prestige, glory, immortality, or some combination of the two. Institutionalists recognize that hegemonic leaders emerge from time to time in the world economy and typically set in motion habits of international cooperation, called "regimes," which consist of "principles, norms, rules and decision-making procedures around which the expectations of international actors converge in given issue areas" (Stephen Krasner, 1983, p. 1). Under British hegemony, the regimes of free trade and the gold standard developed more or less unconsciously. With subsequent American hegemony, a more purposeful process of institution making was undertaken, with agreements at Bretton Woods, on tariffs and trade, the Organization for Economic Cooperation and Development, and the like. Political scientists recognize that regimes are more readily maintained than established since marginal costs are below average costs; as hegemonic periods come to an end with the waning of the leading country's economic vitality, new regimes needed to meet new problems are difficult to create. Cooper (1985) has written of the eighty years it took to create and get functioning the World Health Organization despite the clear benefits to all countries from controlling the spread of disease. And it takes work to maintain regimes; in the absence of infusions of attention and money, they tend in the long run to decay.

I originally suggested that the 1929 depression was allowed to run unchecked because there was no leading country able and willing to take responsibility for crisis management, halting beggar-thy-neighbor policies from 1930, and especially acting as a lender of last resort to prevent the serious run on the Creditanstalt in May 1931 spreading, as it did, to Germany, Britain, Japan, the United States, and ultimately to the gold bloc. Britain, the leading economic power of the nineteenth century, was unable to halt the run; the United States, which might have had the ability, possibly assisted by France, was unwilling. This view has been rejected by one economic historian who holds that the troubles of the interwar period were more deep-seated, and that what was needed was more fundamental therapy than maintaining open markets and providing a lender of last resort, something, that is, akin to the heroic public good after World War II, the Marshall Plan (D. E. Moggridge, 1982). That may have been true, though there is no way I see that the issue can be settled. Leadership at an earlier stage in the 1920's, presumably furnished by the United States with some

cost in foregone receipts on war-debt account, might have resolved the war-debt-reparations-commercial-debt tangle that proved so destabilizing after the 1929 stock market crash. I conclude that the existence of an international lender of last resort made the financial crises of 1825, 1836, 1847, 1866, and 1907 more or less ephemeral, like summer storms, whereas its absence in 1873, 1890, and 1929 produced deep depressions—shortened in the 1890 case by the *deus ex machina* of gold production from the Rand. Again there is room for disagreement.

The point of all this is that after about 1971, the United States, like Britain from about 1890, has shrunk in economic might relative to the world as a whole, and more importantly, has lost the appetite for providing international economic public goods—open markets in times of glut, supplies in times of acute shortage, steady flows of capital to developing countries, international money, coordination of macroeconomic policy and last-resort lending. The contraction of concern from the world to the nation is general, and applies to economists as well as to politicians and the public. In reading recent books on macroeconomic policy by leading governmental economists under both Democratic and Republican administrations, the late Arthur Okun (1981) and Herbert Stein (1984), I have been struck by how little attention the authors paid to international repercussions. The same observation has been made by Ralph Bryant (1980, p. *xviii*) and by the British economist R. C. O. Matthews, reviewing Arjo Klamer's *Conversations with Economists*... (1985, p. 621). There has been a recent upsurge of interest in the international dimension because of the connections among the federal deficit, the exchange rate for the dollar, and the balance-of-payments deficit, but the focus of this interest is almost exclusively on what the connections mean for U.S. interest rates, industrial policy, growth, and wealth. The international impact is largely ignored, bearing out the truth in former German Chancellor Helmut Schmidt's statement that "the United States seems completely unconscious of the economic efforts of its policies on the Alliance" (1984, p. 27).

Some of the discussion of international regimes by political scientists verges on what my teacher, Wesley Clair Mitchell, used to call "implicit theorizing," that is, convenient *ad hoc* theoretical explanations to fit given facts that lack generality. Charles Lipson (1985), for example, suggested that the slippage in U.S. hegemony in the 1970's resulted in a loss of the international public good of secure property rights and therefore in the widespread nationalization of foreign direct investment. He went on to say that the reason less developed countries (*LDC*s) did not default on their debts to bank syndicates was that bank lending was "better institutionalized," "a smaller group," "better protected by legal remedies" (pp. 136, 158, 170). He was surprised that the decline of British hegemony in the interwar period did not result in more *LDC* aggression against foreign property (p. 191), but failed to observe the widespread default on foreign bonds in the 1930's, despite the organization of international finance. In my judgement Keohane exaggerates the efficacy and importance of the international regime in oil that was formed after the first OPEC oil shock of 1973 (see his ch. 10). The crisis caused by the Yom Kippur embargo of the Netherlands was to my mind shockingly mishandled by governments, and the public good of crisis management was left to the private multinational oil companies. The formation of the International Energy Agency was a classic operation in locking the barn door after the horse had been stolen.

Between national self-interest and the provision of international public goods, there is an intermediate position: indifference to both. An interesting contrast has been observed in the 1930's between Britain which forced Argentina into a bilateral payments agreement (the Roca-Runciman Agreement of 1933) in order to take advantage of its monopsony position, and the United States that had a similar opportunity vis-à-vis Brazil but ignored it (Marcelo de Paiva Abreu, 1984).

It is fairly clear from the historical record that economic hegemony runs down in decay—in the British case after 1913 and the United States about 1971—leading Felix

Rohatyn (1984) to say that the American century lasted only twenty years. The Nixon shock of 1973 in cutting off soya bean exports to Japan—a significant harm to an ally for a small gain to this country—was the act of a bad Samaritan. The import surcharge of the same year may have been required to move the dollar out from the position of the nth currency when only $n-1$ countries are free to fix their exchange rates, but it would have been possible to start with the later attempt at cooperation that resulted in the Smithsonian agreement. This is especially true when so much of the case against the 1971 exchange rate was the result of the easy-money policy of the Federal Reserve System under Chairman Arthur Burns, at a time when the Bundesbank was tightening its money market/go-it-alone policies of both banks that flooded the world with dollars.

The present U.S. administration claims to be working for open trade and does fairly well in resisting appeals for protection. The positive push for a Reagan round of trade liberalization in services and agriculture, however, is in pursuit of a national and not an international public good. The regime in capital movements—the World Bank, the regional development banks and that in-last-resort lending orchestrated by the IMF—seems to be working, with bridging loans and an *ad hoc* purchase of oil from Mexico for the U.S. stockpile in 1982 when the IMF finds itself unable to move fast enough. But there are signs of dissension that may spell trouble. The June 1985 bridging loan for Argentina was declined by Germany and Switzerland on the grounds that Argentina had not been sufficiently austere and that its problems were not a threat to the world financial system (*New York Times*, June 15, 1985, p. 1). The Japanese contribution, moreover, was said to have been small, although no figures were given.

What I worry about mostly is exchange policy and macroeconomic coordination. The U.S. Treasury under Donald Regan was committed to the policy of neglect, presumably benign, but in any event ideological. And the commitment to consultative macroeconomic policies in annual summit meetings of seven heads of state has become a shadow play, a dog-and-pony show, a series of photo opportunities—whatever you choose to call them—with ceremony substituted for substance. The 1950's and 1960's, when serious discussions were held at the lowly level of Working Party No. 3 of the O.E.C.D., were superior because the United States and other countries took them seriously.

I am a realist when it comes to regimes. It seems to me that the momentum set in motion by a hegemonic power—if we must use that expression, I prefer to think of leadership or responsibility—runs down pretty quickly unless it is sustained by powerful commitment. The IMF and World Bank were agreed at Bretton Woods largely as a result of the U.S. Treasury: the forms were international, the substance was dictated by a single country (Armand van Dormel, 1978). In the early days of the IMF, Frank Southard told me, if the United States made no proposal, nothing happened. Today the same is true of the European Economic Community: unless Germany and France see eye to eye, which is infrequent, nothing happens. Proposals of great technical appeal from individuals or small countries are not welcomed as the preparatory phases of the World Economic Conference of 1933 demonstrated (see my 1973 book, pp. 210–14). There needs to be positive leadership, backed by resources and a readiness to make some sacrifice in the international interest.

The leadership role is not applauded. When the United States accused the rest of the world of being free riders, Andrew Shonfield countercharged the United States of being a "hard rider," "hustling and bullying the Europeans," "kicking over chairs when it did not get its way" (1976, pp. 86, 88, 102). Furnishing the dollar to the world as international money has brought the United States an accusation of extracting seignorage, although the facts that the dollar is not a monopoly currency and that foreign holdings earn market rates of interest deflect that criticism in sophisticated quarters.

Neglect can verge on sabotage. When the European central banks collaborated to hold the dollar down at the end of February 1985, the conspicuous failure of the United States to participate on a significant scale encouraged speculators not to cover long positions. A former trader for the Federal Reserve

Bank of New York has expressed concern that the habits of central bank cooperation and U.S. official intimacy with the workings of the foreign-exchange market that have been built up over thirty years are being squandered for ideological reasons (Scott Pardee, 1964, p. 2).

Regimes are clearly more attractive in political terms than hegemony, or even than leadership with its overtones of the German *Führerprinzip* or of Italy's *Il Duce*, if not necessarily more so than responsibility. Polycentralism, pluralism, cooperation, equality, partnership, decoupling, self-reliance, and autonomy all have resonance. But it is hard to accept the view, so appealing to the political right, that the path to achieve cooperation is a tit-for-tat strategy, applied in a repetitive game, that teaches the other player or players to cooperate (Robert Axelrod, 1984). As Tibor Scitovsky demonstrated years ago (1937), this path can readily end by wiping out trade altogether. Hierarchical arrangements are being examined by economic theorists studying the organization of firms, but for less cosmic purposes than would be served by political and economic organization of the production of international public goods (Raj Sah and Joseph Stiglitz, 1985).

Minding one's own business—operating in the robust zone of indifference—is a sound rule on trend when macroeconomic variables are more or less stable. To the economist it means reliance on the market to the extent that the conditions for a Pareto optimum solution are broadly met. But the fallacy of composition remains a threat, and one cannot count on the Categorical Imperative. Markets work most of the time, as a positive-sum game in which the gain for one does not imply a loss for another. Experience teaches, however, that crises may arise. When they do, the rule changes from government and public indifference to the production of public goods by leadership or by a standby regime.

Leadership or responsibility limited to crises encounters another problem: how to keep the machinery for handling crises from obsolescence. In crisis one needs forceful and intelligent people, capable of making decisions with speed under pressure. It is sometimes said that the Japanese practice of decision by consensus with ideas coming up from below, makes it hard for that country to discharge in timely fashion the responsibilities of world leadership. In Marcus Goodrich's *Delilah* (1941), the amiable practice of fraternization between a watch officer and enlisted men on the bridge of the destroyer proved dangerous in a typhoon since the men had fallen into the habit of discussing the officer's orders. The paradox is that the attributes needed in crisis tend to atrophy in quiet times; for example in the control room of a Three Mile Island nuclear power plant.

Let me conclude by emphasizing once again my concern that politicians, economists, and political scientists may come to believe that the system should be run at all times by rules, including regimes, not people. Rules are desirable on trend. In crisis the need is for decision. I quote once more the letter of Sir Robert Peel of June 1844 a propos of the Bank Charter Act of that year:

> My Confidence is unshaken that we have taken all the Precautions which Legislation can prudently take against the Recurrence of a pecuniary Crisis. It may occur in spite of our Precautions; and if it be necessary to assume a grave Responsibility, I dare say Men will be found willing to assume such a Responsibility.
>
> [*Parliamentary Papers*, 1857, 1969, p. *xxix*]

REFERENCES

Arrow, Kenneth, J., "Gifts and Exchanges," in Edmund S. Phelps, ed., *Altruism, Morality and Economic Theory*, New York: Russell Sage Foundation, 1975.

Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.

Baltzell, E. Digby, *Puritan Boston and Quaker Philadelphia: Two Protestant Ethics and the Spirit of Class Authority and Leadership*, New York: Free Press, 1979.

Becker, Gary, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

Berry, Brian J. L., "City-Size Distribution and

Economic Development," *Economic Development and Cultural Change*, July 1961, *9*, 573–88.

Braudel, Fernand, *Civilization and Capitalism* (15th-18th Century), Vol. 2, *The Wheels of Commerce*, translated from the French by Sian Reynolds, New York: Harper and Row, 1982.

Breton, Albert, "The Economics of Nationalism" *Journal of Political Economy*, August 1964, *72*, 376–86.

Bryant, Ralph C., *Money and Monetary Policy in Independent Nations*, Washington: The Brookings Institution, 1980.

Calabrese, Guido and Bobbitt, Philip, *Tragic Choices*, New York: W. W. Norton, 1978.

Cooper, Richard N., *The Economics of Interdependence: Economic Policy in the Atlantic Community*, New York: McGraw-Hill, 1968.

______, "World-Wide vs Regional Integration: Is There an Optimal Size of the Integrated Area?," in Fritz Machlup, ed., *Economic Integration: Worldwide, Regional, Sectoral*, New York: Halstead, 1977.

______, "International Economic Cooperation: Is it Desirable? Is it Likely?," *Bulletin*, American Academy of Arts and Sciences, November 1985, *39*, 11–35.

de Paiva Abreu, Marcelo, "Argentina and Brazil During the 1930s: The Impact of British and American Economic Policies," in Rosemary Thorp, ed., *Latin America in the 1930s: The Role of the Periphery in World Crisis*, London: Macmillan, 1984.

Fishkin, James S., *The Limits of Obligation*, New Haven: Yale University Press, 1982.

Fogel, Robert W., *Railroads and American Economic Growth: Essays in Econometric History*, Baltimore: Johns Hopkins Press, 1964.

Goodrich, Marcus, *Delilah*, New York: Farrar & Rinehart, 1941.

Graham, Frank D., *The Theory of International Values*, Princeton: Princeton University Press, 1948.

Harris, Seymour E., *The Economics of New England: Case Study of an Older Area*, Cambridge: Harvard University Press, 1965.

Hicks, John R., *A Theory of Economic History*, London: Oxford University Press, 1969.

Hirschman, Albert O., *Exit, Voice and Loyalty*, Cambridge: Harvard University Press, 1970.

Isard, Caroline and Isard, Walter, "Economic Implications of Aircraft," *Quarterly Journal of Economics*, February 1945, *59*, 145–69.

Johnson, Harry G., "An 'Internationalist' Model," in Walter Adams, ed., *The Brain Drain*, New York: Macmillan, 1968.

Keohane, Robert O., *After Hegemony: Cooperation and Discord in the World Political Economy*, Princeton: Princeton University Press, 1984.

Kindleberger, Charles P., *The World in Depression, 1929–1939*, Berkeley: University of California Press, 1973.

______, "Internationalist and Nationalist Models in the Analysis of the Brain Drain: Progress and Unsolved Problems," *Minerva*, Winter 1977, *15*, 553–61.

______, "Government and International Trade," *Essays in International Finance*, No. 129, International Finance Section, Princeton University, 1978.

______, *Multinational Excursions*, Cambridge, MIT Press, 1984.

______, "Des biens public internationaux en l'absence d'un gouvernment international," in *Croissance, échange et monnaie en économie international, Mélange en l'honneur de Monsieur le Professeur Jean Weiller*, Paris: Economica, 1985.

Knight, Frank H., *The Ethics of Competition and Other Essays*, London: George Allen & Unwin, 1936.

______ **and Merriam, Thornton W.,** *The Economic Order and Religion*, London: Kegan Paul, Trend, Trubner, 1947.

Krasner, Stephen D., *International Regimes*, Ithaca: Cornell University Press, 1983.

Labasse, Jean, *L'espace financier: analyze geographique*, Paris: Colin, 1974.

Lichtenberg, Judith, "National Boundaries and Moral Boundaries," in Peter G. Brown and Henry Shue, eds., *Boundaries: National Autonomy and Its Limits*, Totowa: Rowman and Littlefield, 1981.

Lipson, Charles, *Standing Guard: Protecting Foreign Capital in the Nineteenth and Twentieth Centuries*, Berkeley: University of California Press, 1985.

McCloskey, Donald N., "The Rhetoric of Eco-

nomics," *Journal of Economic Literature*, June 1983, *21*, 481–517.

McKinnon, Ronald I., "Optimum Currency Areas," *American Economic Review*, September 1963, *53*, 717–25.

Matthews, R. C. O., Review of Arjo Klamer, *Conversations with Economists...*, 1983, *Journal of Economic Literature*, June 1985, *23*, 621–22.

Meade, James E., *The Theory of International Economic Policy*, Vol. II, *Trade and Welfare*, New York: Oxford University Press, 1955.

Moggridge, D. E., "Policy in the Crises of 1920 and 1929," in C. P. Kindleberger and J.-P. Laffargue, eds., *Financial Crises: Theory, History and Policy*, Cambridge: Cambridge University Press, 1982.

Mundell, Robert A., "A Theory of Optimum Currency Areas," *American Economic Review*, September 1961, *51*, 657–65.

Nozick, Robert, *Anarchy, State and Utopia*, New York: Basic Books, 1974.

Ohlin, Bertil, *Interregional and International Trade*, Cambridge: Harvard University Press, 1933.

Okun, Arthur M., *Prices and Quantities*, Washington: Brookings Institution, 1981.

Olson, Mancur, *The Rise and Decline of Nations: Economic Growth, Stagflation and Social Rigidities*, New Haven: Yale University Press, 1982.

Pardee, Scott, "The Dollar," address before the Georgetown University Bankers Forum, Washington, D.C., September 22, 1964.

Patinkin, Don, "A 'Nationalist' Model," in Walter Adams, ed., *The Brain Drain*, New York: Macmillan, 1968.

Robertson, Sir Dennis, "What Do Economists Economize?," in R. Leckachman, ed., *National Policy for Economic Welfare at Home and Abroad*, New York: Doubleday, 1955.

Rohatyn, Felix G., *The Twenty-Year Century: Essays on Economics and Public Finance*, New York: Random House, 1984.

Rostow, Walt W., *The Stages of Economic Growth: A Non-Communist Manifesto*, Cambridge: Cambridge University Press, 1960.

Sah, Raaj Kumar and Stiglitz, Joseph E., "Human Fallibility and Economic Organization," *American Economic Review Proceedings*, May 1985, *75*, 292–97.

Scitovsky, Tibor, "A Reconsideration of the Theory of Tariffs," reprinted in AEA *Readings in the Theory of International Trade*, Homewood: Richard D. Irwin, 1949.

Shonfield, Andrew, *International Economic Relations of the Western World*, Vol. I, *Politics and Trade*, New York: Oxford University Press, 1976.

Singer, Peter, "Famine, Affluence and Morality," *Philosophy and Public Affairs*, Spring 1972, *1*, 229–43.

Smith, Adam, *The Theory of Moral Sentiments, or An Essay Toward an Analysis of the Principles by which Men Naturally Judge Concerning the Conduct and Character First of the Neighbours and then of Themselves*, 11th ed., Edinburgh: Bell and Bradfute, 1759; 1808.

———, *An Inquiry into the Nature and Causes of the Wealth of Nations*, Canaan ed., New York: Modern Library, 1776; 1937.

Schmidt, Helmut, "Saving Western Europe," *New York Review of Books*, May 31, 1984, *31*, 25–27.

Stigler, George J., "Economics—The Imperial Science?," mimeo., 1984.

Stein, Herbert, *Presidential Economics: The Making of Economic Policy from Roosevelt to Reagan and Beyond*, New York: Simon and Schuster, 1984.

Van Dormel, Armand, *Bretton Woods: Birth of a Monetary System*, New York: Holmes and Meier, 1978.

Vaubel, Roland, "The Government's Money Monopoly: Externalities or Natural Monopoly?," *Kyklos*, 1984, *27*, 27–57.

Vernon, Raymond, *Sovereignty at Bay*, Cambridge: Harvard University Press, 1971.

———, *Storm Over Multinationals*, Cambridge: Harvard University Press, 1977.

Walzer, Michael, "The Distribution of Membership," in Peter G. Brown and Henry Shue, eds., *Boundaries: National Autonomy and Its Limits*, Totowa: Rowman and Littlefield, 1981.

———, *Spheres of Justice*, New York: Basic Books, 1983.

Parliamentary Papers: Monetary Policy, Commercial Distress, Shannon: Irish University Press, 1957, 1969.

Retail Pricing and Clearance Sales

By EDWARD P. LAZEAR*

A large department store wants to sell a "one-of-a-kind" designer gown. Although the manager has some idea about the price that the gown can command, there is generally some guesswork associated with the process. How should he choose his initial price? If the dress does not sell at that price after the first few weeks on the rack, he has the option of trying a new price. When does he change price and how does the new price relate to the old? How does the decision depend on the characteristics of the gown and on conditions in the clothing market? Is the gown more likely to sell during the first few weeks on the rack, is the pattern of expected transactions smooth over time, or do most sales occur later, when the seller is most frantic about getting rid of the gown?

Firms face a very similar problem when they market a new product that is not unique. Imagine a computer firm that introduces a new model. How should the firm select a time path of prices for the computer, recognizing that the company is not completely certain about the market for the new item? Is it best to start with a high price and lower it over time, or should it do the reverse? Are there any circumstances for which a constant price over time is the appropriate strategy? How is its price contingent on the number of sales made in the first days that the product is on the market? Are the number of transactions likely to be larger at the beginning and then taper off, or might they be smooth over time? How do these patterns vary with the characteristics of the goods and the nature of the buyers? When does the firm announce a "clearance sale," which is an attempt to move merchandise at a price (significantly) below its original price?

This paper provides a simple framework that permits the analysis of these issues. It is an attempt to explain pricing and transaction patterns over time. A number of market phenomena are explained. Among the more interesting ones are

1) Differences in pricing behavior by characteristics of the goods. For example, it is commonly alleged that women's clothes are more expensive than men's clothes, given cost conditions. Also, "designer" items often carry extremely high initial prices, which fall rapidly if the good does not sell. If men's suits do not exhibit such volatile price behavior, what might explain this pattern? What does "fashion" have to do with this and is there an objective definition of fashion that yields predictions?

2) Prices may be more or less variable depending upon the thinness of the market. For some items, say, a $2 million mansion, transactions are relatively rare events. How does the pricing of these infrequently traded items differ from that of goods that turn over often?

3) Do strategies vary with the uniqueness of the good? Designer dresses and Picasso paintings are unique. One and only one item of its exact type is for sale. But there are many copies of a new computer model and the sale of one machine does not preclude the sale of another identical one to another buyer. How does pricing and selling strategy differ in these two cases?

4) Price reduction policies as a function of time on the shelf. Some famous department stores have an announced policy of halving the price of an item for each week that it remains on the floor. Such "bargain basement" behavior can be predicted and the

*University of Chicago, Graduate School of Business, 1101 East 58th Street, Chicago, IL 60637, and National Bureau of Economic Research. Many useful comments were provided by Dennis Carlton, Victor Goldberg, John P. Gould, Robert Hall, Brian Hanessian, Michael Mussa, Peter Pashigian, Sam Peltzman, Melvin Reder, Sherwin Rosen, George Stigler, Jon Strand, and Yoram Weiss. I am especially indebted to Kevin M. Murphy, who caught a number of errors in the first draft. This research was supported by the National Science Foundation.

price-cutting rule can be specified as well. When is a rigid rule of this sort an optimal pricing policy?

The goal is to relate these pricing and selling strategies to underlying, observable characteristics of the market in order to explain the differences. Factors relating to the heterogeneity of the goods, the heterogeneity of buyer preferences, and search costs are discussed.

The idea behind the model is that the ability to sell goods over time allows richer strategies for two reasons. First, if the good does not sell during the first period, the seller still has a chance of selling it during the next period. Second, the outcome of the first period provides the second-period seller with additional information. The amount and nature of that information depends on the characteristics of the market and the number and attributes of the buyers. This can be modeled in a very easy way and all of the questions posed above can be addressed.[1]

A distinction that plays an important role is the one between the expected selling price and the expected revenue associated with a good. The former is the price, conditional on a sale. The latter takes into account that a sale does not always occur. It generally pays to set prices in a way that sometimes leaves the good unsold at the end of the period.

The most important point to bear in mind is that this is a model of "retailing." Retailing, as defined in this paper, describes a selling pattern with an announced price that is maintained for some period of time. The seller agrees implicitly to sell to the first person (or in the case of nonunique goods, to any person) who comes along and is willing to pay that price. The good sells at the stated price. No haggling occurs and auctions, that pit one buyer against another, are not held. Since the analysis and some results bear a close relation to the auction literature, some comparisons are made below. Although the retailing paradigm is taken as given and exogenous, some attempt to explain why retailing is used over other selling schemes is presented in that section as well. Implicit in the fact that goods must be examined to assess value is some *ex post* monopoly power, perhaps created by the imperfect information.

The design of the paper is to start with the simplest model and then to introduce complications as necessary to explain the data. The effects of market competition and strategic behavior of buyers are all considered in turn.

I. Intertemporal vs. Single-Period Pricing

The ability to change price after the first attempt to sell the product fails produces a richer set of strategies and changes the problem facing the firm. To see this, let us begin with the most basic characterization of the firm's pricing problem in a single-period context.

A. *A One-Period Model*

Suppose that the firm will encounter one and only one buyer who is willing to pay V for the good, but no more. The firm does not know V with certainty, but has some prior notion of the density of V denoted $f(V)$ with distribution function $F(V)$.[2] (The prior

[1]Neither the questions nor methodology are entirely novel. The updating procedure used is described in Morris DeGroot (1970) and has been used more recently by Sanford Grossman, Richard Kilstrom, and Leonard Mirman (1977). Another line of literature, Frank Bass (1969), Bass and Alain Bultez (1982), Darral Clarke and Robert Dolan (1984), and Michael Spence (1981) examines some of these problems, but the focus is on cost conditions that change over time, generally tied to some learning by doing. It is my view that the essence of the marketing problem that faces a firm that introduces a new product is selecting a strategy in the face of uncertainty about the demand for its product. The evolution of prices and transactions over time is more likely to reflect learning about the market than learning about producing the product. Both models give declining prices over time, but in the case of cost changes, myopic sellers charge prices that are too high (Bass and Bultez), whereas in the case of learning about demand, myopic sellers choose prices that are too low. In any case, the distinction between this analysis and those that have gone before is that this analysis focuses on demand uncertainty and ignores cost altogether.

[2]It is useful to recognize that $F(V)$ is determined after the retailer has seen the good himself. For example, retail sellers of dresses know that they vary in price

may be based on an examination of the selling prices of similar goods, but for now, its source is unimportant.) The risk-neutral firm's problem is to maximize expected profits, or

$$(1) \qquad \operatorname*{Max}_{R} R[1-F(R)],$$

where R is the price and $1-F(R)$ is the probability that V exceeds R so that a sale is made. For the purposes of expositional simplicity, suppose that the prior on V is uniform between zero and one. Then $F(R)=R$ so that the optimum is at $R=1/2$, yielding expected profits of $1/4$.

An alternative formulation is that the firm has a large number of similar, but not identical items that it wishes to sell. It knows that the distribution of demand prices is given by $f(V)$, but it does not know which items correspond to high values and which to low. An example is a line of dresses that come in different colors or have different trim. *Ex ante*, the seller does not know whether it is the yellow or the red one that has $V=1$. The one-period pricing rule is again, set $R=1/2$.

B. *A Two-Period Model*

Now suppose instead that if the good does not sell during the first period, the seller faces another buyer during the second period who is identical to the one he saw during the first period. The firm now has two chances to sell the good. Furthermore, the failure of the good to sell in period 1 at price R_1 tells the seller something about the V. In this simple case, it implies that $V<R_1$, because if $V>R_1$, the good would have sold.[3]

Using Bayes' Theorem, this implies that the posterior distribution that the firm carries into period 2 is uniform between 0 and R_1, so that $F_2(V)$, the posterior distribution, $=V/R_1$. The choice of R_1 affects the problem in two ways. First, it affects the probability of a sale in period 1. Second, it determines what the firm can infer from no sale. For example, if $R_1=1$, then the fact that the good did not sell is uninformative because the firm was certain that $V<1$ at the outset. Similarly, $R_1=0$ is certain to result in a sale during the first period so that is no learning that occurs with this choice either.

The firm's problem then is to choose R_1 and R_2 where R_2 is the price that the firm tries in period 2, given that the good did not sell in period 1. In this example, the good is unique so that a sale in period 1 eliminates any concern about period 2. That problem can be written as

$$(2) \qquad \operatorname*{Max}_{R_1, R_2} R_1[1-F(R_1)] + R_2[1-F_2(R_2)]F(R_1).$$

The first term is the price charged in period 1 times the probability that the good sells in period 1. The second term is the price charged in period 2 times the probability of a sale in period 2 at that price, given the information from period 1, times the probability that the good does not sell in period 1.

It is instructive to think of this as a dynamic programming problem and to consider the firm's optimal strategy in period 2, given that the good did not sell in period 1 at the price R_1. The firm's problem in period 2 is

$$(3) \qquad \operatorname*{Max}_{R_2} R_2[1-F_2(R_2)].$$

Bayes' Theorem implies that

$$F_2(R_2)=F(R_2)/F(R_1) \qquad \text{for } R_2<R_1$$
$$=1 \qquad \text{otherwise,}$$

so that the first-order condition for (3) can be written

$$(3') \qquad R_2=\left(F(R_1)-F(R_2)\right)/f(R_2).$$

from \$50 to \$10,000. After having examined the good, the seller may know that a particular dress has a V between \$500 and \$1,000, but he does not know the exact value within that range.

[3] This is different from the usual price discrimination problem where the demand curve of the market is known, but no buyer will reveal where he is on that demand curve. That problem is treated by Nancy Stokey (1981) and Jeremy Bulow (1982).

Since $R_2 > 0$, and F is monotonic, $R_1 > R_2$. Price in period 2 is always lower than price in period 1; a clearance sale is held. In the specific case where $V \sim U[0,1]$, the solution is $R_2 = R_1/2$.

This makes obvious intuitive sense. For any given R_1, if the good did not sell during the first period, then the seller can rule out the possibility that $V > R_1$. The distribution that the seller uses in period 2 is uniform between 0 and R_1 so the second period's problem is equivalent to the one facing a firm with only one period to sell and with a prior between 0 and R_1. The solution to that problem is to select $R_2 = (R_1)/2$.

Thus, for any given R_1, if the good remains unsold after one period, the rule is to cut the price in half next period. (This halving is specific to the assumed distribution, of course.) Substitution of $R_2 = (R_1)/2$ into (2) and maximizing with respect to R_1 yields[4]

$$R_1 = 2/3; \qquad R_2 = R_1/2 = 1/3.$$

This illustrates a number of important points. First, prices fall over time. A retailer puts a gown on the market at a high price ($R_1 = 2/3$), hoping that it will sell at that price. If it does not sell, he can revise his price downward during the next period. The reverse pattern would never be optimal, because once the gown sold at the low price, the seller has eliminated the chance that he will get 2/3 for it. (If $V > 2/3$, then it exceeds 1/3 as well so period 2 becomes irrelevant since no gown that will sell in period 2 is ever available after period 1.)[5] Stated alternatively, the gowns in the "prettiest" colors with $V > 2/3$ sell in period 1. The seller must revise downward his opinion of the value of remaining gowns. The distribution at the beginning of period 2 has lower-value gowns than those at the start of period 1 because the best ones have already been picked off.

This abstracts from any investment in brand-name recognition that is associated with charging a lower initial price. For example, new firms frequently charge lower prices than their rivals to induce customers to try the new product. The difference between the observed price and the optimal one as calculated in this problem can be thought of as advertising and is ignored throughout. It also abstracts from contagion or network effects. The value, V, is assumed to be independent of the number of others who have similar items.[6]

Second, the comparison with the one-period solution is interesting. There, the solution was to set price equal to 1/2. Now, because a disappointed seller has another chance, a first-period price that exceeds 1/2 is justified.[7] If he wanted to, he could always select a price of 1/2 during the second period because he still has one chance left. Of course, given what he has learned from period 1, a price of 1/2 is no longer optimal in round 2. So the prices charged straddle the one-period optimum.

Furthermore, expected profits are higher as a result of having a second chance. In the one-period problem, expected profits were 1/4. In the two-period problem, expected profits are 1/3 (substitute $R_1 = 2/3$, $R_2 = 1/3$ into (2)). This is because the expected probability of a sale is higher in the two-period problem. The expected probability that a sale occurs in one of the two periods is

$$(1 - F(R_1)) + (1 - F_2(R_2))F(R_1)$$

or $1 - 2/3 + [1 - (1/3)/(2/3)](2/3) = 2/3$.

In the one-period problem, the expected probability of a sale was $(1 - F(R)) = 1/2$.

[4] The same solution is obtained if (2) is maximized simultaneously choosing R_1 and R_2 because of the time-consistent nature of the problem.

[5] This pricing pattern resembles a Dutch auction. This is discussed in greater detail below.

[6] This is an inaccurate assumption in some cases. For example, the demand for telephones depends on the number of others who can be called.

[7] This result is obtained by Grossman, Kilstrom, and Mirman. Although most of their work focuses on consumer learning, they show that an experimenting monopolist always starts with a higher price than a non-experimenting one.

(The expected selling price is the same in both cases.)[8]

C. *Heterogeneous Consumers and Thin Markets*

The previous problem was made simple because the inference problem was so trivial. If the good did not sell during the first period, the firm knew with certainty that it overpriced the good. In reality, other factors make the inference problem more difficult. Specifically, two factors are important. The first is the number of customers who come into the store during the first period. Intuitively, if only a few customers arrive during the first period, the firm should be less certain about its inference than if a large number examine the good and reject it at price R_1. Second, heterogeneity among consumers may be important. If some consumers are willing to pay V, while others will pay an amount below the firm's reservation price, then the problem is more complicated. The good might not have sold not because the price was too high, but because that period's customers were all of the wrong type.

This can be parameterized as follows. Suppose that in period 1, N "customers" examine the good. Of those, the prior probability P are just "shoppers" whose value of the good is less than the seller's reservation price, and $1-P$ are "buyers" who are willing to pay V. As before, V is unknown to the seller and his goal is to select R_1 and R_2 to maximize profits, given his prior beliefs on V. In what follows, customers refers to the total number of individuals who inspect the good, buyers refers to the subset with value equal to V, and shoppers refers to the subset with value equal to zero. (It is important that a given individual does not know whether he is a buyer or shopper until he has inspected the good. No one who knew that he was a shopper would ever bother to look.)[9]

The problem is similar to the one in (3) except for two points: first, $F_2(V)$ is different here; second, expected sales depend on P. More formally, the seller wants to maximize

$$(4)\quad \max_{R_1, R_2} R_1(\text{Prob. sale in 1}) + R_2(\text{Posterior prob. sale in 2}) \times(\text{Prob. no sale in 1}).$$

Now, the probability of a sale in period 1 is $(1-F(R_1))(1-P^N)$, because the probability that *every* customer is a shopper is P^N so that $1-P^N$ is the probability of encountering at least one buyer. It only requires one buyer to make the sale as long as $R_1 < V$. Similarly, the posterior probability of a sale in period 2 is $(1-F_2(R_2))(1-P^N)$ and the probability of no sale in period 1 is $1-[(1-F(R_1))(1-P^N)]$.

It is now necessary to derive $F_2(V)$. Bayes' Theorem states that the posterior probability is proportional to the probability of the sample, given the parameter, times the prior probability of the parameter. The sample in this case is the observation that no one bought during period 1. For $V < R_1$, the probability of no purchase is 1. For $V > R_1$, there is only one reason why the good did not sell during period 1 and that is that all customers were shoppers. This happens with probability P^N.

It is easy to show that the normalization required to make the integral of the density function equal to 1 is $1/[R_1(1-P^N)+P^N]$ so that the density is given by

$$(5)\quad f_2(V) = \frac{1}{R_1(1-P^N)+P^N} \quad \text{for } V \leq R_1$$

$$= \frac{P^N}{R_1(1-P^N)+P^N} \quad \text{for } V > R_1.$$

[8]Milton Harris and Artur Raviv (1981a) consider the use of different types of pricing mechanisms when demand conditions are uncertain. Their "priority pricing" scheme resembles an intertemporal price decline. They show that such a scheme is optimal when capacity falls short of potential demand. That is the situation that is implicit in this setup because the unique good can be sold to only one of many potential buyers.

[9]This abstracts from shopping for the pure pleasure of it and from any information that might be useful in making future purchases.

Integrating this yields the distribution function

$$(6)\quad F_2(V)=\frac{V}{R_1(1-P^N)+P^N}\quad \text{for } V\le R_1$$

$$=\frac{R_1+P^N(V-R_1)}{R_1(1-P^N)+P^N}\quad \text{for } V>R_1.$$

To obtain the probability of a sale in period 2, $(1-F_2(R_2))$ must be multiplied by $(1-P^N)$ since that is the probability that at least one buyer is encountered in the group of customers.

Substitution of these expressions into (4) yields the following maximization problem:

$$(7)\quad \underset{R_1,R_2}{\text{Max}}\, R_1[1-P^N-(1-P^N)R_1]$$
$$+R_2[P^N+(1-P^N)R_1-R_2](1-P^N).$$

As before, it is instructive to solve this as a dynamic programming problem, deriving the optimal R_2 for any R_1, given that period 2 is reached. This problem is written

$$(8)\quad \underset{R_2}{\text{Max}}\, R_2(1-F_2(R_2))(1-P^N),$$

$$\text{or } \underset{R_2}{\text{Max}}\, R_2\left[1-\frac{R_2}{R_1(1-P^N)+P^N}\right](1-P^N)$$

Differentiating with respect to R_2 and setting the derivative equal to zero yield the optimum R_2 which is given by

$$(9)\quad R_2=\frac{1}{2}[R_1(1-P^N)+P^N].$$

When $P^N=0$, the problem in (7) and (8) reduces to the simpler problem, which is a special case, considered earlier. Indeed, substitution of $P^N=0$ into (9) yields the earlier solution that $R_2=(1/2)R_1$.

Given the solution for R_2 in terms of R_1, (7) can be rewritten as

$$(10)\quad \underset{R_1}{\text{Max}}\, R_1[1-P^N-(1-P^N)R_1]$$
$$+\frac{1}{4}(1-P^N)[P^N+(1-P^N)R_1]^2$$

which yields the solution

$$(11)\quad R_1=\frac{2+P^N(1-P^N)}{4-(1-P^N)^2}.$$

When $P^N=0$ so that the problem reduces to the simple one, the solution is again $R_1=2/3$ and $R_2=(1/2)(2/3)=1/3$. As P^N goes to 1, however, the solution goes to $R_1=1/2$ and from (9), $R_2=1/2$. That is, as P^N goes to 1, prices remain constant over time.

The intuition behind this result is straightforward. When $P^N=0$, all customers are buyers (there are no window shoppers) so the inference problem becomes perfect. If the good is left on the shelf after the first period, it can only be because the good was priced higher than V. Therefore, all $V>R_1$ can be ruled out. But as P^N approaches 1, almost all of the customers are merely shoppers. Thus, little can be inferred from the fact that no one bought the good after the first period. Even if R_1 were less than V, there is a very good chance that the good would still remain on the shelf after one period in this climate of browsers. Under these circumstances, having two consecutive periods is no different from having two independent one-period problems, since nothing is learned from the first period. This implies that the solution to the single-period problem, namely price $=1/2$, applies in both periods.[10]

From (11), it can be shown that

$$(12)\quad \frac{\partial R_1}{\partial P^N}=\frac{-4P^N-(1-P^N)^2}{[4-(1-P^N)^2]^2}<0$$

and that

$$(13)\quad \frac{\partial R_2}{\partial P^N}=\frac{1}{2}\left[1-R_1+(1-P^N)\frac{\partial R_1}{\partial P^N}\right]>0$$

[10] Note that an identical mechanism is at work in the labor market context when trying to infer a worker's product-wage ratio from past transactions. This is the subject of my 1986 paper.

The implication is that when P^N is small, prices start higher and fall more rapidly with time unsold. For P^N close to 1, prices tend to be constant over time. Now, P^N has at least two real-world interpretations.

First, N is the number of customers per period of time. As N increases, P^N gets small so that as N increases, prices start high and fall more rapidly. When there are a lot of customers per period, there is more information contained in the fact that the good did not sell so that the strategy moves toward that used when perfect inference is available. On the other hand, if N is small, less is learned from the fact that the good remains unsold after one round. This implies that the prices of goods in thin markets should start lower and fall less rapidly (relative to the prior distribution) than prices of goods where markets are dense.

Consider, for example, the problem of selling a house. Suppose there are two different types of houses. One is a \$2 million mansion. Such houses turn over very infrequently and there are very few buyers. Another is a \$50,000 high-rise condominium in a building where one of the 300 apartments is sold weekly. The implication of this section is that prices of mansions should be less sensitive to time on the market than prices of condominiums. The reason is that the owner of the mansion cannot infer that his house was overpriced from the fact that it has been on the market for two months without selling. There are very few potential buyers of mansions. But the owner of the apartment can quickly and precisely infer that if the apartment did not sell, it is not because he encountered few buyers, but instead because it was overpriced. Information comes with each genuine buyer and there are fewer of these per period of time in the case of mansions. This suggests that prices of lower-quality goods adjust more rapidly to time on the market during which the goods remains unsold.

The sale of a house does not quite fit this model, since the process is one of haggling over price, rather than strict retailing. Still, the intuition of the example is appealing. Goods for which the markets are thin have more rigid prices; clearance sales are less common. How is thin defined? Since N is the number of customers that any one seller faces, thinness must be defined in some relative sense. Probably the most easily measured aspect of thinness relates to the transactions per unit of time. Consider the house example. If there are 100 houses of the low-priced variety and 5 of the high-priced variety, then in equilibrium, 100 families live in the former and 5 in the latter. Thinness would be the same unless each of the 100 turns over more frequently per unit of time. Suppose that those who live in the low-priced type move twice as often as the high-price residents. Then the number of customers that visit the low-priced houses per unit of time exceeds that at the high-priced houses so an unambiguous measure of thinness can be obtained.

The second interpretation of P^N relates to search cost and information. For a given N, P is the proportion of customers who have a purchase price below the seller's reservation price (in the example above, it was zero). If customers have much information about the good before they inspect it, then few shoppers will show up and all of the customers will be buyers. Consider wholesale vs. retail buyers. It is possible (although not obviously true) that purchasers in the wholesale market have better prior information than those in the retail market. If true, this implies that wholesale prices fall more rapidly with inventory time than retail prices, because the seller at the wholesale level can infer more than the seller at the retail level about his pricing policy. What should be true under any circumstances is, that for a given number of customers, an increase in the proportion of those who do not buy after having examined the good reduces the speed with which prices fall. Both of these variables, the number of customers and the proportion who do not buy after looking, are observable, at least conceptually.

Related to this is the idea that search costs are important in determining the speed with which prices fall as a function of time on the shelf. Consider a good for which search is costly, for example, a piece of land in the middle of Alaska. For a given number of customers, a very small proportion will be

window shoppers. Because inspection is so expensive, most who inspect the good are likely to be buyers rather than shoppers. As such, the seller of that parcel of land can infer a great deal from the decision by any customer not to purchase the land. Thus, the listing price of the land should drop rapidly each time a customer opts against purchase.

Contrast this with a house in the middle of Chicago. The proportion of shoppers to buyers is much higher here because search costs are low. Even individuals who are likely to value the good at zero rather than V may consider taking a look to be certain. Thus, less can be inferred from a given customer's decision not to buy the house. This implies that price is less sensitive to N in Chicago than it is in Alaska. Of course, for a given period of time, N, the number of customers, is likely to be higher for the house in Chicago than for the land in Alaska. This means that prices may fall more rapidly with time even though not with N for the house in Chicago. Both time on the shelf and N are observable.

Additionally, goods for which repeat purchases are made are likely to have informed customers and sellers. The prior on V is tight and P is likely to be small. Recall that the reason that prices fall over time is that learning has taken place during the relevant period. The amount of learning that can occur depends on the dispersion in the prior on V. But if the same good has been sold for a long period of time, the relevant prior is likely to be extremely tight. Thus, little learning occurs and prices remain rigid as a result. Both factors, the length of the time horizon and the amount of dispersion in the prior, are analyzed more rigorously below.[11]

[11] Recently, Paul Milgrom and J. Roberts (1984) have shown that a seller should use price and the level of advertising to signal the quality of the good to the consumer. That will not solve the problem here. They are considering "experience" goods, where repeat sales are important. The idea here is that the good is a search good. Additionally, there is no question about the price that consumers will pay, once the good is identified. Thus, their "free samples" result, that goods are given away at a low price to inform about quality, is not what is at work in this model.

D. *Observable Time Patterns of Price and Quantity*

The theory yields predictions of pricing behavior as a function of three factors: the number of customers, N; the proportion of customers who are shoppers rather than buyers, P; and the firm's beliefs about the market, parameterized through the prior on V. With the exception perhaps of the last of the three, these variables are observable, at least in theory. However, it is likely to prove quite difficult to obtain information on P and N.

Quite aside from data considerations, it is useful to be able to relate price time paths and quantity time paths to some observable characteristics, as well as to each other. The relation of R_1 and R_2 to P and N has already been discussed. Recall that as P^N goes from zero to one, R_1 moves from 2/3 to 1/2, and R_2 moves from 1/3 to 1/2: R_1 falls and R_2 rises so the ratio of R_1 to R_2 falls as P^N increases, that is, as inference becomes more difficult.

The pattern of expected transactions over time is somewhat less intuitive. The probability of a sale in period 1 is

$$\text{(14)} \quad \text{Prob. sale in } 1 = 1 - P^N - (1-P^N)R_1 = (1-P^N)(1-R_1).$$

The (unconditional) probability of a sale in period 2 is

$$\text{Prob(sale in } 2 | R_2, \text{ no sale in } 1) \times \text{Prob(no sale in } 1).$$

This is the second term of (7) without the price R_2 as a scalar. Substituting (9) into this part of (7) yields

$$\text{(15)} \quad \text{Prob. sale in } 2 = \frac{1}{2}\left[P^N + (1-P^N)R_1\right](1-P^N).$$

Division of (14) by (15) gives the ratio of sales in period 1 to those in period 2. That

P 3908

TABLE 1—EXPECTED PRICE AND QUANTITY RELATIONSHIPS

P^N	R_1	R_2	Expected Revenue	Probability of Sale	R_1/R_2
0.0000	0.6667	0.3333	0.3333	0.6667	2.0000
0.0500	0.6610	0.3390	0.3220	0.6441	1.9500
0.1000	0.6552	0.3448	0.3103	0.6207	1.9000
0.1500	0.6491	0.3509	0.2982	0.5965	1.8500
0.2000	0.6429	0.3571	0.2857	0.5714	1.8000
0.2500	0.6364	0.3636	0.2727	0.5455	1.7500
0.3000	0.6296	0.3704	0.2593	0.5185	1.7000
0.3500	0.6226	0.3774	0.2453	0.4906	1.6500
0.4000	0.6154	0.3846	0.2308	0.4615	1.6000
0.4500	0.6078	0.3922	0.2157	0.4314	1.5500
0.5000	0.6000	0.4000	0.2000	0.4000	1.5000
0.5500	0.5918	0.4082	0.1837	0.3673	1.4500
0.6000	0.5833	0.4167	0.1667	0.3333	1.4000
0.6500	0.5745	0.4255	0.1489	0.2979	1.3500
0.7000	0.5652	0.4348	0.1304	0.2609	1.3000
0.7500	0.5556	0.4444	0.1111	0.2222	1.2500
0.8000	0.5455	0.4545	0.0909	0.1818	1.2000
0.8500	0.5349	0.4651	0.0698	0.1395	1.1500
0.9000	0.5238	0.4762	0.0476	0.0952	1.1000
0.9500	0.5122	0.4878	0.0244	0.0488	1.0500
1.0000	0.5000	0.5000	0.0000	0.0000	1.0000

ratio is

$$2(1-R_1)/\left(P^N+(1-P^N)R_1\right).$$

After substituting (11) into this expression, the ratio reduces to 1. That means that the unconditional probability of a sale in period 1 is equal to that for period 2. Expected sales are smooth over time.

Additionally, since expected sales are equal in each period, the probability that the good sells is given by twice the probability that it sells in period 1 or by

$$\text{Prob. of sale} = 2(1-P^N)(1-R_1).$$

This varies with P^N as

$$d(\text{Prob. of sale})/dP^N = -2(1-R_1) \\ -2(1-P^N)\left(\partial R_1/\partial P^N\right).$$

After substitution of (12), it can be shown that

$$\partial(\text{Prob. of sale})/\partial P^N < 0.$$

Table 1 simulates some values.

The relationships illustrated in Table 1 provide empirically testable predictions. As P^N goes from zero to one, the price ratio R_1/R_2 falls. Similarly, as P^N goes from zero to one the probability of an eventual sale falls. This implies that in markets where prices fall rapidly as a function of time on the shelf, the probability that the good will go unsold is relatively low. The prediction in the housing sample is that mansions, for which P^N is high, should have slowly declining prices and should be more likely to be taken off the market after an unsuccessful attempt to sell than inexpensive condos, for which P^N is low.

This implication is not an obvious one. Since R_1/R_2 is high when R_1 is high, the logic implies that for a given prior, goods for which price starts high are actually more likely to sell. The reason is that the high initial price reflects low P^N. It is not a matter of calling out prices randomly. The high initial price is a response to conditions that also imply that a sale is likely.

Note that the expected price at which the good sells is always $1/2$, irrespective of P^N. This follows because (prob. sale in 1)/(prob. sale in 2) $=1$ and because $(R_1+R_2)/2=.5$. This also illustrates the important distinction between the expected price at which a good sells and expected revenue. Although expected price, given a sale is independent of P^N, expected revenue falls with P^N. The probability that the good remains unsold (and is returned to the supplier as scrap) increases with P^N. The point that not all goods are sold and that there is a systematic relationship between pricing and the probability of a sale is fundamental to this analysis. It plays an essential role in reconciling some phenomena described below.

E. *Recapitulation*

The ability to readjust price as a function of past sales provides the firm with a richer strategy set. This is especially important when the firm is more uncertain about the value that consumers attach to the good in question. Not only does intertemporal pricing permit more than one chance to attract buyers, but it also allows the firm to learn about the nature of demand in the market.

An important implication is that prices start high and fall with time on the shelf. The level of initial price and speed with which price falls are positively related to the number of customers that it encounters per period and to the proportion of real buyers in the group. Thin markets have lower initial and more rigid prices.

Competition among firms for customers reduces the number of potential buyers that any one seller encounters. This drives profits to zero, but in the process, alters the optimal pricing rule. Department stores that sell somewhat distinct items will select lower initial prices and lower those prices more slowly when there is competition between stores for buyers. This is true even when the good in question is available at only one store.

The time pattern of transactions tends to be smooth over time. The probability that the good eventually sells is positively related to initial price and the rate of price decline along the optimal price trajectory.

II. Heterogeneous Goods, Fashion, Obsolescence, and Discount Rates

This section builds on the earlier ones to explain how prices vary with factors like product heterogeneity, obsolescence rates, and time discounting. In most of this section, it will be assumed that all customers are buyers, that is, that $P=0$ so that the less complex formulation of the model can be used.

A. *Heterogeneity Among Goods*

Is there any sense to the claim that women's clothes cost more than men's, even for given cost conditions? This is a direct implication of different product heterogeneity across the type of good.

Formally, what this section examines is how dispersion in the prior on V affects pricing policy and the probability that a sale is made. Assume that $P=0$ so that the firm's problem becomes the one in (2) (which is the special case of (4) with $P=0$).

Consider a mean-preserving spread. For expositional convenience, let us be specific. Suppose that the prior on V for, say, men's clothes, is uniform between .5 and 1.5, but for women, it is uniform between 0 and 2. (I ignore the endogeneity of the prior throughout.) The idea is that to the extent that women's clothes take on more variations, it is more difficult to predict the value of any particular item. Both distributions have the same mean value and it would seem that average prices, average revenues, and expected revenues might be the same. This is not the case.

Given the distributions, the prior distribution function for men's clothes is

$$(16a)\quad F(V)=V-1/2 \quad \text{for } 1/2\leq V\leq 3/2$$

and this results in a posterior for any given

R_1 of

$$F_2(V) = (V-1/2)/(R_1-1/2) \quad \text{for } 1/2 \leq V \leq R_1. \tag{16b}$$

Similarly, the prior for women's clothes is

$$F(V) = V/2 \qquad \text{for } 0 \leq V \leq 2 \tag{17a}$$

and this results in a posterior for any given R_1 of

$$F_2(V) = V/R_1 \qquad \text{for } 0 \leq V \leq R_1. \tag{17b}$$

Substitution of (17a), (17b) into (2) yields the solution that the initial price for women's clothing, R_1, equals 4/3 and the period 2 price, R_2, is 2/3. This makes sense since the prior on V is simply a rescaling of the original prior, where solutions were 2/3, 1/3.

Substitution of (16a), (16b) into (2) yields the solution that the initial price of men's clothing, R_1, equals 1 and the period 2 price, R_2, is 1/2. Note that the period 2 price is the lower bound of the posterior (and prior) distributions so that the optimum in this case is to make the sale a certainty in period 2. (Of course, this result is dependent on the shape of the density function.)

Given the prices and the priors, it is obvious that the probability that the woman's garment sells at 4/3 is 1/3 and the (unconditional) probability that it sells at 2/3 is also 1/3. This results in an expected price of 1, and the good is sold 2/3 of the time so expected revenue is 2/3.

For men's clothes, the probability that the garment sells at $R_1 = 1$ is 1/2 and the probability that it sells at 1/2 is also 1/2. The expected price is 3/4 and expected revenue is 3/4.

Although this is only one example, it illustrates a number of important points. First, the more disperse prior results in a higher expected price for a given mean. Thus, women's clothes cost more than men's clothes. Second, because women's clothes remain unsold more often than man's clothes, expected revenues can be lower, even though the price, given a sale, is higher. In competition, firms enter the men's clothing industry until expected revenue is equal across sectors. In the more general model, where $P \neq 0$, this reinforces the result that men's clothes sell for lower prices. The key to this result is that at the optimum prices, more women's clothes remain unsold. The seller either retains the good (as in the case of an unsold house), or wholesales it off.

A similar story might apply to goods that are very new or rapidly changing over time. To the extent that the prior is more diffuse for these goods, their prices should start higher, but fall faster than those on more traditional items. This predicts more variance over time in the prices of new computers than in the prices of standard typewriters. Another reason for high-price variance in the computer market may be the importance of obsolescence. The next section examines that issue.

As an empirical matter, economists who construct price indexes tend to focus on the price, given a sale, and ignore the probability of a transaction. What this points out is that pricing and sale probabilities are linked. For many purposes, when the probability of a sale is less than one, expected revenue per good might be a better metric than expected price, given a sale. The former is more closely related to what the firm generally cares about, even though the latter is what consumers care about.

B. *Fashion, Obsolescence, and Discounting the Future*

Some goods go out of style very quickly whereas others seem to retain their popularity for long periods of time. Again, the example of men's and women's clothes may be relevant. It may be true that men's suits change lapel widths less frequently than women's clothes change style. That phenomenon is assumed exogenous for the purposes of this paper, but it is interesting to know how fashion, or obsolescence as it might be termed in other markets, affects the choice of initial price and the rigidity of prices over time.

This is easily treated in the current framework. Let us think of obsolescence or fashion as taking the following form: during the first

period, the good is worth V, but in the second period it is worth V/K, where $K \geq 1$. The seller still does not know V, but he does know that whatever it is, it will retain only $1/K$ of its worth in period 2.

All that changes is the value that is inserted into the period 2 density function. That is, the individual buys the good in period 2 when $V/K > R_2$, or when $V > KR_2$. During period 1, nothing is changed so the firm's maximization problem in (2) now becomes

$$\text{(18)} \quad \max_{R_1, R_2} R_1[1-F(R_1)] + R_2[1-F_2(KR_2)]F(R_1).$$

Assuming that the prior is uniform between zero and one, the optimum prices are from the first-order conditions

$$\text{(19a)} \quad R_2 = R_1/2K;$$

$$\text{(19b)} \quad R_1 = 2K/(4K-1).$$

For $K=1$, the solutions are identical (as they must be) to those obtained without obsolescence, namely, $R_1 = 2/3$, $R_2 = 1/3$.

What is clear from (19a) is that prices fall faster with time on the shelf when K is large. The reason, of course, is that the seller knows that if the good was worth V in period 1, it is only worth V/K in period 2, so period 2's price adjusts accordingly.

Equally intuitive is that the price in period 1 is lower when K is large. The more obsolete the good becomes, the more anxious is the seller to get rid of it in period 1. As a result, he trades off this sense of urgency against the price that would provide him with the best posterior to carry into period 2.

Stated alternatively, a "classic," defined as a good that does not go out of style, carries a higher initial price, independent of any resale considerations. Its price is less sensitive to inventory than a good that goes out of style rapidly. This is true even for a given set of cost conditions.

Time discounting, although seemingly similar, is somewhat different. The reason is that even though the seller might think of a sale in period 2 at V as worth only V/K in present value, the posterior density function is still on V, not V/K because buyers are willing to pay V in period 2. Thus, the objective function is not (18), but is instead

$$\text{(20)} \quad \max_{R_1, R_2} R_1(1-F(R_1)) + (R_2/(1+r))(1-F_2(R_1))F(R_1).$$

The present value of the period 2 price is $R_2/(1+r)$, but the customer continues to buy the good as long as $V > R_2$. The solution when V is uniform between zero and one is given by

$$\text{(21a)} \quad R_2 = R_1/2;$$

$$\text{(21b)} \quad R_1 = 2(1+r)/(4(1+r)-1).$$

Note that R_2 is $R_1/2$, which differs from (19a). The price does not fall more rapidly when the discount rate is positive. This is as it should be. Given that the firm gets to period 2, the best that it can do is take the information from period 1 (that $V < R_1$) and optimize. Discounting is irrelevant to that decision. This was not true when the good became obsolete in the second period.

But implications about urgency are similar. As r gets large, the firm is anxious to make the sale in period 1, not because the good will become obsolete, but for reasons of time preference. At the extreme, as r goes to infinity, $R_1 = 1/2$. The value of the second period is zero so the firm behaves as it would in the one-period problem, setting $R_1 = 1/2$. But if it does get to period 2, the best policy now is to cut price to $1/4$ because it knows (with certainty) that $V < 1/2$.

Time discounting reduces the initial price, but does not change the rate at which prices fall as a function of time on the shelf. Obsolescence reduces the initial price too, but also increases the rate at which prices fall as a function of time on the shelf.

C. *Longer Horizons*

Two periods have been assumed throughout the analysis. Time discounting was one way to modify that assumption, but it is

useful to consider more directly how a change in time horizon affects pricing strategy. In many respects, this is another way to treat obsolescence, but there is more to it than that.

Consider a firm that has T rather than two periods during which to sell its product. The problem in (2) generalizes to

$$\text{(22)} \quad \max_{R_1, R_2, \ldots, R_T} R_1[1-F(R_1)] + R_2[1-F_2(R_2)]F(R_1) + R_3[1-F_3(R_3)]F(R_2) + \ldots + R_T[1-F_T(R_T)]F(R_{T-1}),$$

where $F_t(V)$ is the posterior after $t-1$ periods. As before, $F(V)$ refers to the prior distribution before period 1. Each term on the right-hand side has as one of its components $F(R_{t-1})$ because this is the probability that the good was not sold before period t. The problem yields a system of recursive first-order conditions given by

$$\text{(23)} \quad \begin{aligned} \partial/\partial R_1 &= 1-2R_1+R_2=0 \\ \partial/\partial R_2 &= R_1-2R_2+R_3=0 \\ \ldots \quad \partial/\partial R_{T-1} &= R_{T-2}-2R_{T-1}+R_T=0 \\ \partial/\partial R_T &= R_{T-1}-2R_T=0. \end{aligned}$$

These yield the general solution that

$$\text{(24a)} \quad R_T=1/(T+1)$$

$$\text{(24b)} \quad R_t=(T-t+1)R_T=\frac{T-t+1}{T+1}.$$

These solutions are quite intuitive. First, as T gets large so that the horizon lengthens, equation (24a) implies that the price in the last period goes to zero. Second, equation (24b) implies that price drops by a smaller amount each period with increases in T. Thus, price changes less rapidly per period. But the initial price is higher as T increases so that a larger total range of prices is covered. As T goes to infinity, $R_1=T/(T+1)$ goes to 1. The price starts at the top and moves down trivially each period until V is hit precisely. As T goes to 1, we are back to the one-period problem and R_1 is 1/2. Efficiency, in that the good is always sold, is guaranteed as T goes to infinity.

Stated simply, as the firm's selling horizon lengthens, initial price is higher and prices fall off less rapidly each period. However, the price in the final period is lower as the time horizon increases. This also implies that the probability that the good sells before the end is reached increases in T because $1-F(R_T)$ increases in T.

The difference between adding periods and merely lengthening the time associated with each period is that learning takes place and a new price can be chosen each period. This comes back to an essential feature of "retailing" as defined in this paper. The price is fixed for a given length of time (which is likely to depend on the number of customers encountered per unit of time). Price changes only occur at the end of that interval. No attempt is made to call out the highest possible price, and lower it until the customer agrees to purchase. There are good reasons for not doing this, and those reasons are discussed below.

D. *Nonunique Goods*

There is another respect in which the time horizon can be lengthened. The situation that many firms face in marketing new products is somewhat different from the one analyzed so far. Above, it was assumed that once the good is sold, there are no others to sell. This is appropriate for a painting or designer dress, but what of a new computer model put out by an established company? The fact that the good sells in the first period does not preclude additional sales in the second period. How should the prices be set under these circumstances?

Again, for simplicity, return to the two-period horizon problem and continue to assume that $P=0$: all customers are buyers. Now, three prices are relevant: the seller must select a price in period 1, R_1; he must choose a price in period 2 given that no purchases were made in period 1, R_2; and he

must choose a price in period 2 given that at least one purchase was made in period 1, $\tilde{R}_2$. (Under the assumptions about consumer homogeneity, knowing the exact number of items sold provides no additional information.)[12]

Normalize such that one item is available for sale in period 1 and N_2 are available in period 2. If no sale occurs in period 1, then N_2+1 are available. N_2 may be greater or less than 1. The preceding analysis of unique goods is merely a special case, with $N_2=0$. The firm's maximization can be written as

$$(25)\qquad \underset{R_1,R_2,\tilde{R}_2}{\text{Max}}\; R_1(1-F(R_1)) + (N_2+1)R_2(1-F_2(R_2))F(R_1) + N_2\tilde{R}_2(1-\tilde{F}_2(R_2))(1-F(R_1)).$$

It is especially revealing to treat this as a dynamic program and to examine what happens if the good sells in period 1.

The second-period problem is

$$(26)\qquad \underset{\tilde{R}_2}{\text{Max}}\, N_2\tilde{R}_2(1-\tilde{F}_2(R_2)).$$

Again, using Bayes' Theorem

$$\tilde{f}_2(V)=0 \qquad \text{for } V\le R_1$$
$$=f(V)/(1-F(R_1)) \quad \text{for } V>R_1$$

so
$$\tilde{F}_2(V)=0 \qquad \text{for } V\le R_1$$
$$=(F(V)-F(R_1))/(1-F(R_1)) \quad \text{for } V>R_1.$$

The maximization in (26) becomes

$$(27a)\qquad \underset{\tilde{R}_2}{\text{Max}}\, N_2\tilde{R}_2 \qquad \text{if } \tilde{R}_2<R_1$$

$$(27b)\qquad \underset{\tilde{R}_2}{\text{Max}}\, \frac{N_2\tilde{R}_2(1-F(R_2))}{1-F(R_1)} \qquad \text{if } \tilde{R}_2\ge R_1$$

Branch (27a) is always increasing in $\tilde{R}_2$ so if the solution is on this branch, it is at $\tilde{R}_2=R_1$. If $\tilde{R}_2\ge R_1$, then the first-order condition for (27b) is relevant:

$$\frac{d}{d\tilde{R}_2}=\frac{N_2}{1-F(R_1)}(1-F(\tilde{R}_2) - \tilde{R}_2 f(\tilde{R}_2))=0$$

or
$$\tilde{R}_2=(1-F(\tilde{R}_2))/f(\tilde{R}_2).$$

This solution is identical to that of the one-period problem. But the optimal price in the one-period problem can never exceed R_1, so the corner is relevant here, too. Thus, the solution is $\tilde{R}_2=R_1$.

This implies that price in the second period never rises, even if the good sells during the first period. The reason is that the part of the distribution below R_1 is irrelevant anyway, so knowing that $V>R_1$ does not change the decision on the optimal price.

Given that $\tilde{R}_2=R_1$, and using the definition of $\tilde{F}_2(V)$, equation (26) can be rewritten as

$$\underset{R_1,R_2}{\text{Max}}\, R_1(1-F(R_1)) + (N_2+1)R_2(1-F_2(R_2))F(R_1) + N_2R_1(1-F(R_1))$$

or
$$\underset{R_1,R_2}{\text{Max}}\,(N_2+1)R_1(1-F(R_1)) + (N_2+1)R_2(1-F_2(R_2))F(R_1).$$

Since the scalar (N_2+1) is irrelevant, this problem is identical to the one in (2), where goods were assumed to be unique. Thus, all results already derived hold even in the case of nonunique goods. In this example, $R_1=2/3$; $R_2=1/3$; $\tilde{R}_2=2/3$. If some sales are made in period 1, then price is held at R_1. If no sales are made, then price is halved.[13]

[12] Secondhand markets are ignored.

[13] The solution that price never rises after a successful period depends critically on two assumptions. First, there are no contagion or network effects that shift

E. *Spoiling the Market and Nondurable Goods*

The result that the price following a successful period 1 never falls hinges on the assumption that demanders who find the price too high in period 1, return for another look in period 2. (If no sale occurs, there are N_2+1 buyers in period 2.) This is an inappropriate assumption in two obvious cases. The first is that buyers lose interest when they find that $R_1>V$. This may be rational when buyers do not know the firm's prior so that they cannot forecast its price-cutting behavior. The second is that the good is nondurable. For example, a hotel room that was vacant on Saturday night cannot be stored and sold again on Sunday.

Under these circumstances, the maximization problem is

$$\begin{aligned}\max_{R_1, R_2, \tilde{R}_2} \; & R_1(1-F(R_1)) \\ & + R_2N_2(1-F_2(R_2))F(R_1) \\ & + \tilde{R}_2N_2(1-\tilde{F}_2(R_2))(1-F(R_1)).\end{aligned}$$

Using the results of the previous section, this can be written as

$$\begin{aligned}\max_{R_1, R_2} \; & (N_2+1)R_1(1-F(R_1)) \\ & + R_2N_2(1-F_2(R_2))F(R_1).\end{aligned}$$

The first-order conditions imply that at the optimum, $\tilde{R}_2=R_1$; $R_2=R_1/2$; and $R_1=(2N_2+2)/(3N_2+4)$.

If $N_2=1$ so that demand is constant over time, $R_1=4/7$, instead of $2/3$ as obtained before. The reason is that losing a sale in period 1 is now more costly since the market is spoiled for that buyer, so the firm selects a lower first-period price. The learning effect, which is still present, is offset to some extent by the desire to avoid having first-period buyers walk away without buying. It is not offset completely because $R_1>1/2$, so this is not the same as consecutive one-period problems. Even though the good is not storable, the information derived from period 1 is, so sellers of nondurables do not behave myopically. As N_2 gets large, R_1 approaches $2/3$ because the lost sales in period 1 are trivial, relative to revenue generated in period 2. The information effect dominates.

IV. Consumer Behavior

There are a number of aspects of consumer behavior that are worth considering. I start by analyzing strategic play by purchasers and link this analysis to the auction literature.

A. *Strategic Considerations*

Consumers may know the firm's pricing policy and in particular, that $R_2<R_1$. Does it pay for a consumer in period 1 to wait for period 2, knowing that by doing so he may be able to purchase the good at a lower price? The decision depends on the number of rival customers.

Suppose that a buyer has located a gown in period 1 that she values at $V>R_1$. If she buys the dress, she earns rent $=V-R_1$. Alternatively, she can wait until period 2 hoping that no others will get there first. The more potential customers there are in the market, the lower is the probability that the gown will remain on the rack into the next period. If the buyer passes up the gown this time, there are $N-1$ other customers who might beat her to it next period. Therefore, the expected rent from waiting is $(V-R_2)/N$. She waits if[14]

$$V-R_1<(V-R_2)/N,$$

or if $\quad R_1(1-1/2N)>V(1-1/N)$

since $R_1=2R_2$.

As N goes to infinity, the left side goes to R_1 and the right to V. So the consumer waits

demand in period 2 relative to period 1. Second, the group of buyers is homogeneous in the assessment of V.

[14] This ignores one-period bargaining considerations.

if $R_1 > V$. But $R_1 > V$ precludes buying the good in period 1 anyway, so for sufficiently large N, strategic behavior is not an issue. There are too many others around who can beat this customer to it. She buys it when she finds it.[15] The argument is reinforced if some of this period's buyers might obtain the good first.

On the other hand, as N goes to 1 it is certain that the consumer behaves strategically because $R_1/2 > 0$. The consumer is sure to get it next period since she has no competition so she might as well wait for a lower price. Thus, strategic behavior is not an issue when there is a large number of potential buyers, but may be important when only a few individuals are even potentially interested in the good.[16]

If goods are not unique, then without time preference, strategic behavior results in an equilibrium that is identical to that of the one-period problem. The reason is that all buyers gain if no sales are made in period 1. Since the good is not unique, all are satisfied in period 2 at a lower price ($R_2 < R_1$). No buyer has any incentive to purchase in period 1. Of course, if the seller knows this, then he can infer nothing from the fact that no one bought in period 1. As such, his problem is like the one-period problem so the solution is $R_t = 1/2$. Given that solution, no strategic waiting occurs.

The original solution (2/3, 1/3, 2/3) is restored if some buyers have high time preferences for the good, or if the seller can induce one of the buyers to reveal the information. (A calculator that produces services over time provides more utility, the sooner it is acquired.) Then at least some buyers have an incentive to deviate from the waiting strategy. The initial "no waiting" equilibrium is restored when some buyers have sufficiently high time preference to buy in period 1. This requires $V - R_1 > (V - R_2)/(1+r)$ where r is the discount rate. For r sufficiently large, a sale in period 1 is guaranteed when $V > R_1$. Then the solution reverts to setting $\tilde{R}_2 = R_1$ so that strategic waiting is not an issue. Since price does not fall over time, nothing is gained by waiting.

B. *Auctions and Stochastic Arrival of Customers*

One way to deal with uncertainty about consumer demand is to hold an auction.[17] In fact, the solution to the basic problem is in many respects simply a Dutch auction, where the price begins high and continues to fall until a purchaser declares that he is willing to buy at that price. In the case where the time horizon is long, so that the reduction in price is small at each period, and where N is small so that consumers may behave strategically with respect to waiting time, the analysis is that of a traditional Dutch auction.

There are two major differences between this analysis and the one that pertains to the standard Dutch auction. The first is one of emphasis. This analysis for the most part assumes that N is large and consequently ignores most strategic behavior by consumers.[18] It focuses instead on the rule that the *seller* uses to choose the optimal size of

[15] The argument here is a special case of the more general one made by Robert Wilson (1977). He shows that as the number of bidders gets large, a sealed-bid auction results in bids that are almost certain to be equal to the true reservation value. As will be pointed out below, the declining price retail policy is like a Dutch auction, which is equivalent to a sealed-bid auction in fundamental respects. As such, the Wilson result is relevant in this context.

[16] This assumes that the seller ignores the strategic behavior of consumers. This assumption is troublesome, and without it, a pure strategy equilibrium may not exist.

[17] There is a large literature on auctions starting with William Vickrey (1961). He explores the allocative and profit implications of a number of different kinds of auctions, including the second price and Dutch auction. A number of recent papers have characterized the conditions under which various types of auctions are efficient and profit maximizing. Among those are Gerard Butters (1975), R. Engelbrecht-Wiggans (1980), Harris-Raviv (1981b), Roger Myerson (1981), and Milgrom and R. J. Weber (1982).

[18] Most of the auction literature focuses on strategic behavior by consumers in selecting a bid. An early example of this kind of analysis is Wilson (1967), who examines what happens in an auction when one and only one party is informed about the value of the good.

the step as a function of the number of bidders and their types (shoppers or buyers), and of the number of periods in the horizon. Recall that the number of periods depends on the cost of changing price because a period is defined as that time during which price does not change. If there were no cost to changing price, then the period notion would be dispensed with and the seller would change price each time a customer examined the good. Then R would move with N only and time would be irrelevant.

The second important difference is related. In a Dutch auction, there is no reason to alter the size of the step once the process is in motion. That is, once the seller selects an amount by which to lower price each time, no new information appears until a buyer agrees to purchase the good, at which point it is too late to use that information. That is not true in the retailer context, nor is it true in this model. Although we have assumed throughout that N is fixed, there is nothing in the setup of the problem that precludes a stochastic N. In fact, for a given choice of R_1, the optimal R_2 is given by equation (9), reproduced here:

$$R_2 = \frac{1}{2}\left[R_1(1-P^N)+P^N\right].$$

If N here is interpreted as the realization of N in period 1, then (9) still holds as the optimum R_2 (because the second-period N does not enter). Thus, the retailer can alter his choice of R_2, that is, change the size of the step in the Dutch auction, after having observed something from the first period. Of course, the ability to do so changes the choice of R_1 because that problem now involves an integral over all possible realizations of N.[19]

[19]The problem in (7) then becomes

$$\underset{R_1, R_2}{\text{Max}}\ E\left\{R_1\left[1-P^{\tilde{N}_1}-(1-P^{\tilde{N}_1})R_1\right]\right.$$
$$\left.+R_2\left[P^{\tilde{N}_1}+(1-P^{\tilde{N}_1})R_1-R_2\right](1-P^{\tilde{N}_2})\right\}.$$

where $\tilde{N}_t$ is the stochastic number of buyers in period t.

But the point is that the retail pricing policy has an additional instrument that is useful in all cases where N is stochastic. That instrument is the ability to select the size of the step after obtaining some information.

There are at least two related reasons why a seller might choose a strict retail pricing rule over some form of auction or haggling. The first, already mentioned, is that an auction may encourage strategic behavior on the part of consumers that is absent when retail pricing is used. For example, consider confronting every potential buyer with the maximum possible V and lowering price very rapidly until the consumer agrees to purchase. Obviously, the consumer would wait until $P = 0$, knowing that the considerations that prevented strategic waiting in the last section are not relevant with this type of pricing behavior. The retail pricing method, where R_1 is chosen and fixed in advance, discourages strategic behavior by consumers when the number of customers is sufficiently large. This suggests that retailing is more likely to be used when there are a large number of anonymous buyers.

The second, and perhaps more compelling, reason why large, impersonal stores might prefer retailing to some form of haggling has to do with delegation of authority. Even if no agency problems exist, it is not unreasonable to believe that the management of a department store would not trust price setting to low-paid retail clerks. Since buyers can always refuse to buy if the price is too high, but can purchase if the price is too low, bad price setting can result in losses to the firm even if those prices are only randomly too high or too low. To avoid this adverse-selection problem, the firm may decide to have its experts announce a rigid price (or price rule) that is posted. No haggling is permitted because the clerk who represents the store may not be good at it. Agency problems reinforce this result.

This suggests that haggling is more likely to occur when the owner (who is presumably the high-quality price setter) is also the sales agent. Mom-and-pop stores are more likely to bargain with their customers over price than are large department stores, which use retail pricing almost exclusively.

There are some more subtle elements, special to this setup, that are not generally part of the Dutch auction. First, with a Dutch auction, all bidders are present at the same time. Equivalently,[20] each may submit a binding sealed bid. In the case of the former, it is necessary that customers examine the good at the same time. The story in this paper allows individuals to arrive at different times during each period. The first one to arrive who will pay price R_1 in period 1 (or R_2 if it goes to period 2) gets the good. The Dutch auction imposes the cost that a common meeting time must be found. Retail pricing does not impose that cost because consumers can choose their own shopping time.

Sealed bids do not force a common meeting time, but they do create a waiting period between the time that the bid is made and the winner is determined. This, too, is costly. For example, a woman bids on a dress for the "ball" and then sees another one before she learns whether she made the winning bid on the first—buying the second dress might leave her with two, but failure to purchase might result in her having none.

V. Empirical Thoughts

The purpose of this theory is to provide some empirical implications on pricing and transactions behavior as a function of some observable parameters. There are a substantial number of predictions about pricing and time on the shelf as a function of the number of customers per period, the type of customers (shoppers or buyers), the time horizon, interest rates, the durability of the good, and the shape of the prior.

With the exception perhaps of the last, all of these have observable analogues. The number of customers and general thinness of the market can be proxied by the turnover rate for the good. For example, houses that turn over more rapidly are sold in markets with higher N. The type of customer, P, can be measured by the proportion of individuals who examine a good relative to the number that actually buy it. The time horizon relates to the maximum number of times that a price is changed before the good is taken off the floor; for example, bargain basements eventually give up on some goods. After how many price reductions does this occur?

Additionally, some state contingent behavior has been predicted. For example, when goods are nonunique, the price in period 2 that follows a successful period 1 differs in a specific way from the price that follows an unsuccessful period 1. This relationship is observable and can be tested. Similarly, the time distribution of transactions is related to the initial price and to the speed with which price falls over time. Again, both are observable. Finally, more uncertainty in the prior implies a higher initial price with more rapid decline. It also implies that at the optimal prices, more goods are left unsold when the prior is diffuse. (Fewer men's suits are left unsold than women's dresses.) A relationship between initial price, rate of price decline, and proportion of unsold goods is predicted, and all are observable.

VI. Summary and Conclusion

Sellers must gauge the market any time they attempt to sell a new item. Their attempt to do so and to learn from experience leads to pricing and selling behavior that varies in predictable ways with some observable characteristics of the market.

The major theme is that prices start high and fall as a function of time on the shelf. The speed of that fall and the initial price itself increase as the number of customers per unit of time increases, as the proportion of customers who are "genuine buyers" as opposed to "window shoppers" increases, and as prior uncertainty about the value of the good increases. The optimum price path implies that in the case considered, the probability of making a sale is constant over time.

A number of additional predictions are obtained. First, diffuse priors may result in higher initial prices and more rapid fall, but also in higher average price and more goods left unsold. This might explain why women's clothes carry a higher average price than

[20] See John Riley and William Samuelson (1981).

men's, but are of lower average "quality," even in a competitive market.

Second, goods that become obsolete more rapidly or are more susceptible to fashion exhibit lower initial prices as well as prices that fall more rapidly with time on the shelf. Positive discount rates have a similar effect on initial price, but not the same effect on the rate of price fall.

Third, nonuniqueness of the good does not alter the solution. A successful first period is followed by no change in the price, whereas an unsuccessful first period results in the same price reduction as is warranted when goods are unique.

Fourth, for nonstorable goods, or when spoiling the market is an issue, the price-reduction policy is the same but the initial price is lower than for storable goods. It is higher than the price that a myopic seller of nondurables would charge because even though the good is not storable, the information derived from period 1 is of value.

Fifth, rigid price reduction policies used by bargain basements are predicted under certain circumstances.

Finally, the paper examines strategic behavior by consumers and the relationship between a retailer's pricing policy and Dutch auctions.

REFERENCES

Bass, Frank, "A New Product Growth Model for Consumer Durables," *Management Science*, January 1969, *15*, 215–27.

______ **and Bultez, Alain,** "A Note on Optimal Strategic Pricing of Technological Innovations," *Marketing Science*, Fall 1982, *1*, 371–438.

Bulow, Jeremy, "Durable Goods Monopolists," *Journal of Political Economy*, April 1982, *90*, 314–32.

Butters, Gerald, "Equilibrium Price Distributions and the Economics of Information," unpublished doctoral dissertation, University of Chicago, 1975.

Clarke, Darral G. and Dolan, Robert J., "A Simulation Analysis of Alternative Pricing Strategies for Dynamic Environments," *Journal of Business*, January 1984, *57*, S179–99.

DeGroot, Morris H., *Optimal Statistical Decisions*, New York: McGraw-Hill, 1970.

Engelbrecht-Wiggans, R., "Auctions and Bidding Models: A Survey," *Management Science*, February 1980, *26*, 119–42.

Grossman, Sanford, Kilstrom, Richard and Mirman, Leonard, "A Bayesian Approach to the Production of Information and Learning by Doing," *Review of Economic Studies*, October 1977, *44*, 533–47.

Harris, Milton, and Raviv, Artur, (1981a) "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review*, June 1981, *71*, 347–65.

______ **and** ______, (1981b) "Allocative Mechanisms and the Design of Auctions," *Econometrica*, November 1981, *49*, 1477–99.

Lazear, Edward P., "Raids and Offer-Matching," *Research in Labor Economics*, *8*, 1986 forthcoming.

Milgrom, Paul and Roberts, J., "Price and Advertising Signals of Product Quality," mimeo., Yale and Stanford universities, May 1984.

______ **and Weber, R. J.,** "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, *50*, 1089–122.

Myerson, Roger, "Optimal Auction Design," *Mathematics of Operations Research*, February 1981, *6*, 58–73.

Riley, John and Samuelson, William, "Optimal Auctions," *American Economic Review*, June 1981, *71*, 381–92.

Spence, A. Michael, "The Learning Curve and Competition," *Bell Journal of Economics*, Spring 1981, *12*, 49–70.

Stokey, Nancy, "Rational Expectations and Durable Goods Pricing," *Bell Journal of Economics*, Spring 1981, *12*, 112–28.

Vickrey, William, "Counterspeculation, Auctions and Competitive Sealed Tenders," *Journal of Finance*, March 1961, *16*, 8–37.

Wilson, Robert, "Competitive Bidding with Asymmetric Information," *Management Science*, July 1967, *13*, 816–20.

______, "A Bidding Model of Perfect Competition," *Review of Economic Studies*, October 1977, *44*, 511–18.

Informational Implications of Interest Rate Rules

By MICHAEL DOTSEY AND ROBERT G. KING*

Policy discussions in central banks have long centered on the selection of an appropriate level of the interest rate. Recently, this focus has been translated into a search for an optimal interest rate rule. This paper examines the effects of interest rate rules in a rational expectations macro model that incorporates flexible prices and informational frictions. Specifically, we consider a policy of actively targeting the nominal interest rate, which we define as adjusting its expected level to economic conditions. We find that this class of interest rate rules alters the information content of market prices. Therefore, they alter the magnitude of fluctuations in real activity through the expectational channels described by King (1982).[1] Further, this type of interest rate targeting is shown to be equivalent to a money stock rule with feedback.

The seminal work on this topic is William Poole's analysis (1970) of optimal monetary policy instruments in the context of a textbook Keynesian model, which yielded two major results. First, when the state of the economy is known by the monetary authority, money stock and interest rate policies are equivalent, so that optimal demand management can be achieved by either means. Second, when the central bank cannot fully observe the contemporaneous state of the economy, policymakers should employ the new information contained in the nominal interest rate to counteract the output effects of unobservable real and nominal shocks. Following this line, a policy of "leaning against" interest rate movements (i.e., a policy having a positive interest elasticity of contemporaneous money supply in response to interest rate shocks) is typically desirable. In addition to providing the standard framework for analysis of monetary policy, Poole's 1970 work also hinted at a positive analysis of monetary authorities' observed concern with interest rate smoothing.[2]

In rational expectations models with flexible prices and informational frictions,[3] the implications of Poole's policy alternatives are dramatically altered. In these models, an interest rate rule cannot be arbitrarily postulated—since this leads to nominal indeterminancy—but rather requires the specification of an underlying money supply rule, which serves as a nominal anchor to the system.[4] That is, an even more fundamental equivalence obtains in our rational expectations macro model than in Poole's analysis.

*Federal Reserve Bank of Richmond, Research Department, Richmond VA 23261, and Department of Economics, University of Rochester, Rochester, NY 14627, respectively. We have benefited from the comments of Marvin Goodfriend and presentation of this paper at the Econometric Society meetings in December 1983. The National Science Foundation supported King's participation in this research. The views expressed in this paper are not necessarily those of the NSF, the NBER, or the Federal Reserve Bank of Richmond.

[1]King stresses that differential information on the part of economic agents is a necessary condition for monetary policy to affect the information content of prices in models which perceived monetary policy is neutral toward real activity. An additional necessary condition is that differentially informed agents observe a common price that imperfectly aggregates their information.

[2]See, for example, the discussion of interest rate smoothing in the context of a descriptive analysis of monetary policy provided by Poole (1975). Marvin Goodfriend (1983) offers a positive theory of monetary policy that incorporates an interest rate smoothing objective.

[3]Robert Lucas (1972, 1973) provided initial models that stressed the importance of informational frictions for aggregate supply theory. More recent treatments incorporate economywide bond markets (Robert Barro, 1980; Sanford Grossman and Lawrence Weiss, 1982; and King, 1983) so that discussion of instrument choice becomes interesting.

[4]Thomas Sargent and Neil Wallace (1975) introduce the indeterminacy of the price level that obtains with an arbitrary interest rate rule under rational expectations. Bennett McCallum (1981b, 1983) discusses some alternative ways of resolving this indeterminacy, which all amount to specification of a nominal anchor for the system by a determinate path for the money supply.

Further, the distribution of real activity is invariant to the sort of contemporaneous policy response discussed by Poole, because private agents efficiently utilize the information contained in the nominal interest rate, which is not affected by a known policy of leaning against surprise movements in the interest rate.[5]

The expectational channels that provide the impetus for the nonneutrality of our interest rate rules are quite different from the standard feedback mechanisms analyzed in prerational expectations literature. In particular, we conclude that interest rate targeting can have an impact on the variability of real activity because it affects the information content of prices. But, in contrast to Poole, our analysis provides no reason to prefer an active interest rate policy to a money stock rule with feedback to economic conditions.

When the monetary authority must operate without information about the current aggregate state, so that such an optimal money supply rule or interest rate targeting scheme no longer is feasible, then one must compare two alternative nonactivist policies, a strict money stock rule and an unconditional interest rate peg. Although such an interest rate peg destroys information, it also absorbs money demand disturbances and eliminates money supply shocks. Thus, in a conclusion reminiscent of Poole (1970, Section V), either a strict money stock rule or interest rate peg may be optimal when there are information constraints on the monetary authority.

In Section I, we lay out the simple rational expectations model with flexible prices and informational frictions that we employ in our analysis of policy. In Section II, we discuss the solution of the model with the details presented in the Appendix. In Section III, we consider how monetary policy potentially influences expectation formation and, hence, real activity in our model, with particular attention paid to the informational implications of alternative interest rate rules. In Section IV we examine alternative monetary policies. Section V is a brief summary and presents our conclusions based on this paper and related efforts.

I. The Model

In this paper, we employ a simple macro model to demonstrate a set of results concerning interest rates and informational efficiency. But many of these results also hold in other more complicated models that have flexible prices and information frictions (such as King, 1983, and our earlier 1983 paper).

There are two elements in the model economy that are particularly important for our subsequent policy analysis. First, current commodity supply and demand depend on agents' rational expectations about the real rate of return as in Robert Lucas (1972) and Robert Barro (1980). Second, the economy is populated by two types of agents, who are differentiated by their endowment of information.[6] Specifically, a fraction λ of agents is accurately informed about the contemporaneous state of the economy. The remaining fraction $(1-\lambda)$ is assumed to know only the current values of prices, but not the underlying shocks that determine these prices.[7] Taken together, these two elements dictate limitations on the role for monetary policy previously discussed by Barro (1976) and King (1982). That is, unless the monetary authority has superior information, current feedback to the state of the economy has no real effects, as perceived money growth is

[5]See King (1983), our article (1983), and Mathew Canzoneri et al. (1983) for alternative discussions of this irrelevance result, which requires that agents observe nominal interest rates and that unanticipated but accurately perceived money growth has no real effects.

[6]By viewing the information structure as exogenous, we abstract from equilibrium in the information market as considered by Mark Edwards (1981). The endogenously determined fraction of informed traders would plausibly respond to policy, an effect which is not considered here.

[7]Our basic results do not require that one group is fully informed or, even, that some agents are better informed than others. The key assumption is that agents are differentially informed (see King, 1982, and our earlier paper). The assumption of fully informed agents yields, however, the simplest analytical solutions.

neutral. With differential information, prospective feedback can alter the information content of market prices and, hence, the distribution of real activity.

A. *Commodity Demand and Supply*

Supply and demand at a given date t are aggregates of the actions of informed and uninformed agents. In common with other intertemporal substitution models of business fluctuations, commodity supply and demand depend on the real rate of return expected by market participants. In our log-linear model the real rate of return between t and $t+1$ is $r_t = P_t + R_t - P_{t+1}$, where P_t is the logarithm of the price level at date t and R_t is the level of the nominal interest rate. Informed agents form their rational expectations using a complete information set containing all shocks to the system in date t and earlier periods, which we denote I_t. Uninformed agents are limited to current information about the price level and the interest rate, an information set which we denote $U_t = \{P_t, R_t, I_{t-1}\}$.

Commodity supply and demand are specified as

$$(1) \quad y_t^s = (1-\lambda)\alpha^s Er_t|U_t + \lambda\alpha^s Er_t|I_t - \lambda\beta^s Eg_t|I_t - (1-\lambda)\beta^s Eg_t|U_t + \theta^s g_t$$

$$(2) \quad y_t^d = -(1-\lambda)\alpha^d Er_t|U_t - \lambda\alpha^d Er_t|I_t + \lambda\beta^d Eg_t|I_t + (1-\lambda)\beta^d Eg_t|U_t + \theta^d g_t + \varepsilon_t.$$

In addition to the intertemporal substitution influences of the rate of return, commodity supply and demand also depend on some current disturbances g_t and ε_t. We think of g_t as being an unobservable level of government spending, which has direct supply effects ($\theta^s g_t$) via productivity and demand effects ($\theta^d g_t$); that is, government's purchase of goods less substitution influences on private commodity demand. The coefficients β^s, β^d reflect wealth effects on commodity supply and demand. For a more detailed description of the effect of government purchases on supply and demand decisions, see Barro (1976 or 1980). The term ε_t is a disturbance to private commodity demand.

Commodity market clearing requires that the real rate of return expected by uninformed agents satisfy

$$(3) \quad Er_t|U_t = \lambda(Ep_{t+1}|I_t - Ep_{t+1}|U_t) + ((\theta+\beta)/\theta)g_t + \frac{(1-\lambda)\beta}{\alpha}(g_t - Eg_t|U_t) + \frac{1}{\alpha}\varepsilon_t,$$

where we have defined the composite parameters $\theta = \theta^d - \theta^s$, $\alpha = \alpha^s + \alpha^d$, and $\beta = \beta^s + \beta^d$. (The derivation also employs the fact that $Eg_t|I_t = g_t$.) Substituting this expression into the supply schedule, we obtain the commodity market-clearing value of output

$$(4) \quad y_t = y_t^* + (1-\lambda)(H/\alpha)(g_t - Eg_t|U_t),$$

where the full-information level of output (y_t^*) is

$$(5) \quad y_t^* = (G/\alpha)g_t + (\alpha^s/\alpha)\varepsilon_t.$$

In these expressions, we have used the composite parameters G and H, defined as $G = \alpha^s(\theta - \beta) + \alpha(\theta^s + \beta^s)$ and $H = \alpha^s\beta^d - \beta^s\alpha^d$, which are treated as positive in our analysis.[8]

[8] Given the results of Barro and King (1984), a few words concerning these assumptions are in order. Barro and King show that in models where agents' preferences are time separable and where commodities are nonstorable, the parameter G is positive under standard assumptions, but H is zero. Therefore, output will never deviate from its full-information value regardless of the degree of confusion about the actual value g_t. In order for misperceptions of money and real disturbances to have an effect on output, one must do away with either the time separability or perishable commodity assumptions. However, the resulting models would be extremely complicated. We therefore view the assumption of H greater than zero as a convenient device for analyzing the consequences of misperceptions on output. In the context of the subsequent analysis, all we really desire is a reduced-form solution in which misperceptions of nominal quantities have real effects. Since the underlying structural model plays only a limited role in the results obtained, the above assumptions have no qualitative effect on our results and significantly simplify the analysis.

B. *Money Demand and Supply*

The demand for money is taken to have the semilogarithmic form used by Thomas Sargent and Neil Wallace, and by Barro (1980),

$$(6)\quad M_t^d = P_t + \delta y_t - \gamma R_t - kv_t - (1-k)v_{t-1},$$

where M_t^d is the logarithm of money demand and v_t is an aggregate velocity shock with persisting effects on the demand for money.[9]

Following our discussion above, we specify that the money supply rule involves both responses to interest rate surprises and to the prior state of the economy.

$$(7)\quad M_t^s = \tilde{M}_t + \psi(R_t - ER_t|I_{t-1}) + f_t + m_t.$$

In this expression, $\tilde{M}_t = M_0 + nt$ is the long-run growth path of money and m_t is a random shock to the money supply. Responses to interest rate shocks are captured by the term $\psi(R_t - ER_t|I_{t-1})$, with an interest rate peg obtaining when ψ is driven to infinity. We restrict attention in specifying feedback (f_t) from velocity shocks and past errors in monetary control to the money stock; that is,

$$(8)\quad f_t = f_m m_{t-1} + f_v v_{t-1}.$$

Based on prior work of Bennett McCallum (1981a,b; 1983) and ourselves (1983), we know that one might alternatively view the authority as selecting an interest rate rule. We discuss this possibility in greater detail later in the paper.

II. Rational Expectations Solution

Commodity market and monetary equilibrium yields two equations that link the price level and the nominal interest rate:

$$\begin{aligned}(9)\quad P_t = & -R_t + Ep_{t+1}|U_t \\ & + \lambda(Ep_{t+1}|I_t - Ep_{t+1}|U_t) \\ & + ((\theta+\beta)/\alpha)g_t + ((1-\lambda)\beta/\alpha) \\ & \times (g_t - Eg_t|U_t) + (1/\alpha)\varepsilon_t;\end{aligned}$$

$$\begin{aligned}(10)\quad R_t = & (1/(\gamma+\psi))\{P_t + \delta[(G/\alpha)g_t \\ & + (1-\lambda)(H/\alpha)(g_t - Eg_t|U_t) \\ & + (\alpha^s/\alpha)\varepsilon_t] - kv_t - (1-k)v_{t-1} \\ & - \tilde{M}_t + \psi ER_t|I_{t-1} - f_m m_{t-1} \\ & - f_v v_{t-1} - m_t\}.\end{aligned}$$

Given the structure of the economy, the following undetermined coefficients solutions can be postulated:

$$\begin{aligned}(11)\quad R_t = & \phi_0 + \phi_1\tilde{M}_t + \phi_2 m_{t-1} + \phi_3 v_{t-1} \\ & + \phi_4 m_t + \phi_5 v_t + \phi_6 g_t + \phi_7 \varepsilon_t,\end{aligned}$$

$$\begin{aligned}(12)\quad P_t = & \pi_0 + \pi_1\tilde{M}_t + \pi_2 m_{t-1} + \pi_3 v_{t-1} \\ & + \pi_4 m_t + \pi_5 v_t + \pi_6 g_t + \pi_7 \varepsilon_t.\end{aligned}$$

The details of the solution method are spelled out in the Appendix. As is frequently the case in this class of rational expectations models, one can first and most simply solve for the part of the equilibrium solution that involves the dependences of prices and interest rates on elements of I_{t-1}. These solutions have the following intuitive form

$$\begin{aligned}(13)\quad ER_t|I_{t-1} = & \phi_0 + \phi_1\tilde{M}_t + \phi_2 m_{t-1} + \phi_3 v_{t-1} \\ = & n - (f_m/(1+\gamma))m_{t-1} \\ & - ((f_v + (1-k))/(1+\gamma))v_{t-1},\end{aligned}$$

$$\begin{aligned}(14)\quad EP_t|I_{t-1} = & \pi_0 + \pi_1\tilde{M}_t + \pi_2 m_{t-1} + \pi_3 v_{t-1} \\ = & \gamma n + \tilde{M}_t + (f_m/(1+\gamma))m_{t-1} \\ & + ((f_v + (1-k))/(1+\gamma))v_{t-1}.\end{aligned}$$

That is, the nominal interest rate has an unconditional mean n equal to the trend rate of monetary expansion (the real rate of interest is zero due to the absence of constant terms in (1) and (2)). The price level depends one-to-one on the trend money stock ($\tilde{M}_t$)

[9] The first-order moving average parameterization of money-demand disturbances was chosen for analytical tractability rather than empirical realism.

and depends positively on the rate of monetary expansion (via the inverse effect on cash balances of expected inflation, the intensity of which is governed by γ). Temporarily high values of the money stock $(f_m m_{t-1})$ raise the price level and lower the nominal interest rate via an expected deflation effect. Similarly, the net influence of v_{t-1} involves its own serial correlation (governed by k) and policy influence (governed by f_v), but otherwise works like a temporary money supply disturbance.

A. *Expectations of Uninformed Agents*

Following Lucas (1973) and Barro (1976, 1980), we view uninformed agents as extracting information from available signals contained in prices and interest rates. Given that departures of output from its full-information value are induced solely by $g_t - Eg_t|U_t$, we focus on this expectation, which takes the form

$$(15) \qquad Eg_t|U_t = b_P s_{Pt} + b_R s_{Rt},$$

where s_{Pt} and s_{Rt} are signals contained in the price level and interest rate.[10]

By observing the price level as expressed in (9) and the nominal interest rate given by (10) agents receive the following effective signals,

$$(16) \quad \lambda\pi_2 m_t + \lambda\pi_3 v_t + ((\theta - \lambda\beta)/\alpha) g_t + (1/\alpha)\varepsilon_t \equiv s_{Pt};$$

$$(17) \quad (1/(\gamma+\psi))\{-m_t - kv_t + (((G+(1-\lambda)H)/\alpha))g_t + (\alpha^s/\alpha)\varepsilon_t)\} \equiv (1/(\gamma+\psi))s_{Rt}.$$

[10]The conventional way to derive these signals, as in Lucas and Barro, would be to use the undetermined coefficients representation (11), so that the signal provided by the nominal interest rate would be $\phi_4 m_t + \phi_5 v_t + \phi_6 g_t + \phi_7 \varepsilon_t$. Here, we employ an alternative solution strategy developed by Zvi Hercowitz (1980) which culls "effective signals" from prices and interest rates by using the fact that in equilibrium, agents know the influence of their own expectations on prices. This strategy frequently leads to sharper intuition and more readily obtainable solutions.

There are two important facts to notice about these signals. First, the price level signal is influenced by the expectations of informed agents, $EP_{t+1}|I_t = \pi_2 m_t + \pi_3 v_t$, so long as λ is not zero. Second, the information provided by the interest rate is not altered by any finite value of the policy parameter ψ, as agents can simply "rescale" and learn the same linear combination of fundamental disturbances. But, when ψ is infinite, the interest rate signal is lost.

In interpreting our subsequent analysis, it will be useful to discuss expectation information in the case where there are no nominal shocks. Then, with the two signals s^*_{Pt} and s^*_{Rt} (the asterisk indicating the absence of nominal disturbances) depending only on the two underlying shocks g_t and ε_t, agents can accurately infer the value of g_t. Thus, in this case,

$$(18) \qquad Eg_t|U_t = b^*_P s^*_{Pt} + b^*_R s^*_{Rt} = g_t,$$

where b^*_P and b^*_R are population regression coefficients.

III. Monetary Policies and Expectations

In this section, we explore the effects of some alternative monetary policies on expectations and, hence, on output.

A Strict Money Stock Rule. Under this policy, there is neither contemporaneous response to interest rates ($\psi = 0$) nor feedback from unpredictable changes in money demand ($f_v = 0$). Further, all policy errors (m_t) are eliminated ($f_m = 0$).

Contemporaneous Response to Interest Rates. As discussed previously, Poole (1970) puts forward the hypothesis that contemporaneous money supply response to interest rates can stabilize economic fluctuations. This is depicted by a nonzero, finite value of the parameter ψ in equation (7).

An Interest Rate Peg. An interest rate peg is the limiting case of a contemporaneous response to interest rates (i.e., $\psi = \infty$). Specifically, the monetary authority supplies any quantity of nominal balances demanded at a specified level, which need not be constant from period to period, but must be consistent with the long-run system proper-

ties.[11] Under such a peg, money supply disturbances are eliminated from the system and the signal s_{Rt} destroyed.

An Interest Rate Target. We define a policy of interest rate targeting as adjusting its expected level $(ER_t|I_{t-1})$ to economic conditions, but permitting surprise movements in the interest rate to occur in response to shocks. McCallum (1981a, b; 1983) has taught us that such a policy of interest rate targeting is feasible under rational expectations, as long as (i) the monetary authority provides a nominal anchor to the system (such as $\tilde{M}_t$ in our analysis), and (ii) the authority selects among a class of feasible interest rate rules. In our context, feasible interest rate targets with responses to m_{t-1} and v_{t-1} take the form

$$R_t^T = n + \tau_m m_{t-1} + \tau_v v_{t-1}. \tag{19}$$

Clearly, there is an equivalence between the specification of money supply feedback parameters (f_m, f_v) and specification of the interest rate target parameters (τ_m, τ_v).[12] That is, altering f_m or f_v implies a change in the conditional distribution of the nominal interest rate, since from (13), $ER_t|I_{t-1} = n - (f_m/(1+\gamma))m_{t-1}((f_v + (1-k))/(1+\gamma))v_{t-1}$, and is therefore equivalent to moving the targeted level of the interest rate.

Given the various monetary policies, which are specific cases of the money supply rule (7), we wish to compare the characteristics of each and decide which rule is optimal. Before doing so, we will present a general discussion of the effects of policy on the information content of prices and formulate a means of comparing policies. Placing our specific problem in such a general setting will suggest other contexts in which our results are likely to arise.

A. *Information and Policy*

In this section, we want to distinguish between two different ways in which policy can alter the informational state of the economy, building a foundation for our comparison of the various monetary policies discussed above.

To make our general analysis comparable to the specific problem addressed in the paper, we focus on a case in which economic agents are forming rational perceptions about a single variable x_t, which is itself not directly observable. For this purpose, agents have a vector of information variables or signals $S_t = \langle s_{1t}, \dots, s_{qt} \rangle$. If x_t, S_t are jointly normally distributed—conditionally on the information set A_{t-1}—then it follows that

$$Ex_t|S_t,\ A_{t-1} = \mu_x + b_{xS}(S_t - \mu_S), \tag{20}$$

where $\mu_x = E(x_t|A_{t-1})$, $\mu_S = E(S_t|A_{t-1})$ and $b_{xS} = \sigma_{xS}\Sigma_{SS}^{-1}$ for $\sigma_{xS} = Ex_tS_t|A_{t-1}$ and $\sigma_{SS} = ES_tS_t'|A_{t-1}$. The conditional variance of x_t given S_t is

$$\sigma_{xx} = \sigma_{x'S}\Sigma_{SS}^{-1}\sigma_{xS}, \tag{21}$$

where $\sigma_{xx} = Ex_t^2|A_{t-1}$.

Throughout our discussion, we use the magnitude of the variance of x_t, conditional on a specified information set, as our measure of the "informational state" of the agent or economy under study. That is, when there is a lower value of

$$E(x_t - Ex_t|A_t)^2|A_t, \tag{22}$$

where A_t is the current information set, we say that there is a better informational state.

As a result of our econometrics training and practice, much of our intuition about the effects of information is obtained from the sort of "regression" model outlined above. In the subsequent discussion, we use that intuition to discuss the two basic ways that policy can alter the information state of the economy.

[11] For a more detailed discussion of the determinacy properties of various pegs see McCallum (1981a, b; 1983) and our earlier paper. In general, the resolution of indeterminacy involves the specification of an underlying money stock rule.

[12] One can also view a more restricted form of an interest rate target as $f(ER_t|I_{t-1} - ER_t)$. In the present case where only the past history of velocity shocks is important, this type of response would be equivalent to feedback on a velocity shock.

The first and simplest way is to alter the list of signals while holding the covariance structure fixed. Then it is easy to determine the effect on the informational state. That is, let the information set available to agents include the covariance structure and a proper subset of the signal vector S_t; then, from elementary statistical theory, we know that the conditional variance of x_t increases, lowering the informational state. That is, with the covariance structure fixed, a reduction in the number of signals worsens the informational state. Viewing (18) as a population regression, this accords with our basic intuition, in that fewer independent variables lead to a larger population variance.

The second way that policy can effect the information state is by altering the covariance structure of $x_t, s_{1t}, \ldots, s_{qt}$. In this case there is an ambiguous effect on the informational state. One needs to specify the precise nature of the alteration in covariance structure to determine the effect on the conditional variance of a prediction with fixed number of signals, $E(x_t - Ex_t|A_t)^2|A_t$.

In models where agents form rational perceptions about state variables that are not directly observable, almost all policy interventions affect the covariance structure, while some policy interventions affect the number of signals.

IV. Alternative Monetary Policies

Here we consider some alternative feasible monetary policies. We start by considering an optimal feedback policy or, equivalently an optimal interest rate target. Then, we compare a money stock rule to an unconditional interest rate peg.

A. *Optimal Monetary Policy*

Following the preceeding discussion, we define the optimal interest rate or monetary policy as the one that produces the highest informational state of the economy. Specifically, the optimal policy is that which produces the lowest conditional variance of g_t given the information set of uninformed agents. Thus, our objective will be to find the policy that minimizes $E(g_t - Eg_t|U_t)^2|U_t$.

In our discussion in Section III, we saw that agents could correctly infer g_t from s^*_{Pt} and s^*_{Rt}. It is now easy to show that a feedback policy can effectively alter the information content of prices allowing agents to infer g_t even in the presence of nominal shocks. The conditional expectation of g_t is now given by

$$(23)\quad Eg_t|U_t = b_P s_{Pt} + b_R s_{Rt} = b_P(s^*_{Pt} + \lambda\pi_2 m_t + \lambda\pi_3 v_t) + b_R(s^*_{Rt} - m_t - k_{vt}).$$

If feedback parameters f_m and f_v are set so that $b_P\lambda\pi_2 - b_R = 0$ and $b_P\lambda\pi_3 - b_R k v_t = 0$ then equation (23) will be identical to equation (18) with $b_P = b^*_P$ and $b_R = b^*_R$. That is, agents will be able to infer g_t accurately with optimal feedback. In essence, optimal feedback is able to negate the contaminating influence of nominal shocks in the price and interest rate signals, allowing for a full-information solution. We stress that this can only occur in the presence of differential information ($\lambda \neq 0$), as in the analysis of King (1982) and Lawrence Weiss (1980).

The optimal values of f_m and f_v are $f^*_m = ((1+\gamma)/\gamma\delta\alpha^s)$ and $f^*_v = ((k(1+\gamma))/\gamma\delta\alpha^s) - (1-k)$ (see the Appendix for derivations). Further, an optimal interest rate targeting scheme with $\tau^*_m = \phi^*_2$ and $\tau^*_v = \phi^*_3$ results in a full-information solution, where ϕ^*_2 and ϕ^*_3 are the values of ϕ_2 and ϕ_3 when f_m and f_v are at their optimal levels. This reflects the fundamental equivalence of these two policies, as in Poole's 1970 analysis. However, in contrast to Poole, the optimal policy does not involve responses to unpredictable movements in the current interest rate, nor does it involve any consideration of relative variances.

B. *An Interest Rate Peg vs. a Money Stock Rule*

Suppose, however, that the monetary authority is unable to follow a feedback rule of the type discussed, perhaps because information on lagged shocks is unavailable. Since the information embodied in P_t and R_t is

unaffected by finite values of ψ, we find a policy that contemporaneously responds to interest rates produces the same solution as a pure money stock rule. As our earlier paper stresses, this result occurs because agents are both rational and observe the interest rate. Therefore, movements in money caused by responses to interest rates are perfectly perceived and have no consequence for output or the information content of prices.

This leaves the monetary authority with a choice between a strict money stock rule and a policy of pegging the interest rate at its unconditional expected value n. These alternative policies are the only feasible ones given the limited information possessed by the monetary authority and correspond to the passive policies of Poole (1970). The comparison of these two policies in terms of the informational state produced by each is nontrivial, since the peg alters (i) the number of signals, and (ii) the covariance structure of the model when compared with the money stock rule. That is, under a peg, the interest rate is no longer a signal. Furthermore, under a peg, money supply disturbances no longer arise and velocity shocks are completely absorbed by changes in the money stock.

To compare these two policies, we find it useful to decompose the conditional variance of g_t in a way that highlights the signalling role of the interest rate. We start by calculating the expectation of g_t conditional on observation of the price level. Then we revise this expectation using the information contained in the interest rate. (One can view this procedure as arising because agents receive information sequentially.) Formally,

$$(24)\quad Eg_t|s_{Pt}, s_{Rt} = Eg_t|s_{Pt} + a(s_{Pt} - Es_{Pt}|s_{Rt}),$$

where $Eg_t|s_{Pt} = [\sigma_{gP}/\sigma_{PP}]s_{Pt}$,

$$ER_t|s_{Pt} = [\sigma_{RP}/\sigma_{PP}]s_{Pt}$$

$$a = \{\sigma_{gR} - \sigma_{gP}\sigma_{RP}/\sigma_{PP}\}/\{\sigma_{RR} - \sigma_{RP}^2/\sigma_{PP}\},$$

$$\sigma_{PP} = \mathrm{var}(s_{Pt}), \qquad \sigma_{RR} = \mathrm{var}(s_{Rt}),$$

$$\sigma_{gP} = \mathrm{cov}(g_t, s_{Pt}), \qquad \sigma_{RP} = \mathrm{cov}(s_{Rt}, s_{Pt}).$$

The second term on the right-hand side reflects the extent to which agents revise their expectations based on information contained in the interest rate. The analogous formula for the variance is

$$(25)\ E[(g_t - Eg_t/U_t)^2] = (\sigma_{gg} - (\sigma_{gP}^2/\sigma_{PP})) - a(\sigma_{gR} - (\sigma_{gP}\sigma_{RP}/\sigma_{PP})).$$

The second term on the right-hand side of (25) is nonnegative. Thus, the information contained in S_{Rt} cannot worsen the variance of the prediction error.

By examining (24) and (25) it is easy to see that the peg destroys a signal, wiping out the second right-hand term in (24) and (25). From this standpoint the peg reduces the informational state of the economy. However, it also reduces the variance of the price signal, which no longer contains any nominal disturbances, thereby lowering the first right-hand side term in (25) and improving the informational state of the economy.

In general, neither the peg nor the money stock rule is dominant. Instead the comparison depends on relative magnitudes of the variance of the disturbances and on the parameters of the model, a conclusion which is reminiscent of Poole (1970, Section V). For example, as the variance of the disturbance to private commodity demand is smaller ($\sigma_\varepsilon^2 \to 0$) pegging the interest rate would allow the price level to accurately communicate g_t and the full-information solution would obtain. However, as $\sigma_\varepsilon^2 \to \infty$, the added information contained in the interest rate would imply a dominance of the money stock rule.

V. Summary and Conclusions

This paper explores the informational implications of interest rate policies in rational expectations models with flexible prices and informational frictions. Overall, our analysis suggests that there is little to be gained by discussing monetary policy in terms of interest rate rules. Yet, it does not rule out real effects of systematic actions framed in those terms.

More specifically, we have three major findings of our analysis. First, as in some

earlier analyses, we find that contemporaneous response of the variety discussed by Poole (1970) is not an important determinant of real activity because it simply represents a perceived monetary action and hence, is neutral. Second, and more importantly, we find that interest rate targets can affect the distribution of real activity via the information content of prices, but only in a manner identical to the effects of money supply feedback rules. Finally, we find that either a strict money stock rule or interest rate peg may be a dominant policy, when the monetary authority operates in a situation of incomplete information.

This third finding suggests a potentially rewarding avenue of further research. Suppose that only a subset of the aggregate state of the economy is observable by the monetary authority and the private sector. Then, it appears that there would be a nontrivial choice between a money supply feedback rule, based on observable state elements, and a policy of pegging the interest rate at a level conditional on the observable state elements. This latter policy appears to characterize the actual behavior of the monetary authority in the United States and could potentially be a desirable response in a situation of incomplete information.

APPENDIX

As shown in the text, equilibrium in the goods and money market gives the following two conditions:

$$\text{(A1)} \quad P_t = -R_t + EP_{t+1}|U_t + \lambda(EP_{t+1}|I_t - EP_{t+1}|U_t) + ((\theta-\beta)/\alpha)g_t + (((1-\lambda)\beta)/\alpha)(g_t - Eg_t|U_t) + (1/\alpha)\varepsilon_t;$$

$$\text{(A2)} \quad R_t = (1/(\gamma+\psi))\{P_t + \delta[(G/\alpha)g_t + (1-\lambda)(H/\alpha) \times (g_t - Eg_t|U_t) + (\alpha^s/\alpha)\varepsilon_t] - kv_t - (1-k)v_{t-1} - \tilde{M}_t + \psi ER_t|I_{t-1} - f_m m_{t-1} - f_v v_{t-1} - m_t\}.$$

Using (A1) and (A2) and the undetermined coefficients solutions postulated in (11) and (12), we can solve for the undetermined coefficients attached to the elements of I_{t-1}, namely $\tilde{M}_t$, m_{t-1} and v_{t-1}. These solutions are

$$\pi_0 = \gamma; \quad \phi_0 = 0; \quad \pi_1 = 1; \quad \phi_1 = 0;$$

$$\text{(A3)} \quad \pi_2 = (f_m/(1+\gamma)); \qquad \phi_2 = -(f_m/(1+\gamma))$$

$$\pi_3 = ((f_v(1-k))/(1+\gamma))$$

$$\phi_3 = -((f_v + (1-k))/(1+\gamma)),$$

and are independent of the policy parameter ψ, which simply controls policy responses to shocks.

To study incomplete information, we note that the price level and interest rates are equivalent to observing signals

$$\text{(A4)} \quad s_{Pt} = \lambda\pi_2 m_t + \lambda\pi_3 v_t + ((\theta-\lambda\beta)/\alpha)g_t + (1/\alpha)\varepsilon_t;$$

$$\text{(A5)} \quad s_{Rt} = -m_t - kv_t + \delta((G+(1-\lambda)H)/\alpha)g_t + \delta(\alpha^s/\alpha)\varepsilon_t.$$

The solution for b_p^* and b_R^* in the regression $Eg_t|U_t = b_P^* s'_{Pt} + b_R^* s'_{Rt}$ are straight forward to obtain, by employing $g_t = Eg_t|U_t$ and $s'_{Pt} = ((\theta-\lambda\beta)/\alpha)g_t + (1/\alpha)\varepsilon_t$ and $s'_{Rt} = \delta((G+(1-\lambda)H)/\alpha)g_t + (\alpha^s/\alpha)\varepsilon_t$. We find that $b_P^* = \alpha\alpha^s/D$ and $b_R^* = -\alpha/\delta D$ where $D = (\theta-\lambda\beta)\alpha^s - (G+(1-\lambda)H)$. The solution for f_m^* and f_v^* are found by using the expressions for π_2 and π_3 and solving the restrictions $b_P^*\lambda\pi_2 - b_R^* = 0$ and $b_P^*\lambda\pi_3 - kb_R^* = 0$.

REFERENCES

Barro, Robert J., "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, January 1976, *2*, 1–32.

______, "A Capital Market in an Equilibrium Business Cycle Model," *Economet-*

rica, September 1980, *48*, 1393–1417.

______ and King, Robert G., "Time-Separable Preferences and Intertemporal Substitution Models of Business Fluctuations," *Quarterly Journal of Economics*, November 1984, *99*, 817–38.

Canzoneri, Mathew, Henderson, Dale and Rogoff, Kenneth, "The Information Content of Interest Rates and the Effectiveness of Monetary Policy Rules," *Quarterly Journal of Economics*, November 1983, *98*, 545–66.

Dotsey, Michael, and King, Robert G., "Monetary Instruments and Policy Rules in a Rational Expectations Environment," *Journal of Monetary Economics*, September 1983, *12*, 357–82.

Edwards, Mark S., "Informational Equilibrium in a Monetary Theory of the Business Cycle," working paper, University of Rochester, May 1981.

Goodfriend, Marvin S., "Rational Expectations, Interest Rates Smoothing, and the Optimality of a Non-Trend-Stationary Money Supply Rule," working paper, Federal Reserve Bank of Richmond, July 1983.

Grossman, Sanford J. and Weiss, Lawrence, "Heterogeneous Information and the Theory of the Business Cycle," *Journal of Political Economy*, August 1982, *90*, 699–727.

Hercowitz, Zvi, "Money and the Dispersion of Relative Prices," unpublished doctoral dissertation, University of Rochester, 1980.

King, Robert G., "Monetary Policy and the Information Content of Prices," *Journal of Political Economy*, April 1982, *90*, 247–49.

______, "Interest Rates, Aggregate Information and Monetary Policy," *Journal of Monetary Economics*, August 1983, *12*, 199–234.

Lucas, Robert E., Jr., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

McCallum, Bennett T., (1981a) "Rational Expectations and Macroeconomic Stabilization Policy: An Overview," *Journal of Money, Credit and Banking*, November 1981, *13*, 716–46.

______, (1981b) "Price Level Determinacy with an Interest Rate Policy Rule and Rational Expectations," *Journal of Monetary Economics*, November 1981, *8*, 319–29.

______, "Some Issues Concerning Interest Rate Pegging, Price Level Determinacy, and the Real Bill Doctrine," preliminary working paper, Carnegie-Mellon University, September 1983.

Poole, William W., "Optimal Choice of Monetary Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, *84*, 197–216.

______, "The Making of Monetary Policy: Description and Analysis," *Economic Inquiry*, June 1975, *13*, 253–65.

Sargent, Thomas J. and Wallace, Neil, "Rational Expectations, the Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, *83*, 241–54.

Weiss, Lawrence, "The Role of Active Monetary Policy in a Rational Expectations Model," *Journal of Political Economy*, April 1980, *88*, 221–33.

______, "Interest Rate Policies and Informational Efficiency," Cowles Foundation Discussion Paper No 589, Yale University, April 1981.

Woglom, Geoffrey, "Rational Expectations and Monetary Policy in a Simple Macroeconomic Model," *Quarterly Journal of Economics*, February 1979, *93*, 91–105.

An Operational Measure of Liquidity

By STEVEN A. LIPPMAN AND JOHN J. MCCALL*

What is liquidity? Kenneth Boulding says that "liquidity is a *quality* of assets which... is not a very clear or easily measurable concept" (1955, p. 310). According to John Maynard Keynes:

> There is, clearly, no absolute standard of "liquidity" but merely a scale of liquidity—a varying premium of which account has to be taken...in estimating the comparative attractions of holding different forms of wealth. The conception of what contributes to "liquidity" is a partly vague one, changing from time to time and depending on social practice and institutions. [1936, p. 240]

Similarly, Helen Makower and Jacob Marschak observe that

> ..."liquidity" has so often been used to cover all properties of money indiscriminately that it seems better not to use it for any of the separate properties of money. We thus resign ourselves to giving up "liquidity" as a measurable concept: it is, like the price level, a bundle of measurable properties. [1938, p. 284]

However, they also note that the term liquidity suggests "the fact that money is easily transformable (on the market) into other assets and is thus an effective instrument for manoeuvring" (p. 284). Closely related is the notion of liquidity due to Jack Hirshleifer who said that liquidity is "an asset's capability over time of being realized in the form of funds available for immediate consumption or reinvestment—proximately in the form of money" (1968, p. 1).[1] The notion of liquidity presented here most closely resembles Hirshleifer's.

The purpose of this paper is to present a precise definition of liquidity in terms of its most important characteristic—the time until an asset is exchanged for money. We then show that this definition is compatible with several other useful notions of liquidity.

Whereas academic economists do not possess a definition of liquidity as a measurable concept (though they do mention an assortment of its attributes), other workers in the area casually respond that liquidity is the length of time it takes to sell an asset (i.e., convert into cash); thus cash is considered the most liquid asset, while stocks listed on the NYSE are viewed as more liquid than collectibles, precious metals, jewels, real estate, and capital goods. The problem with this view of liquidity is the lack of precision and casual reference to "the" length of time it takes to convert the asset into cash.

This length of time is a function of a number of factors including frequency of offers (i.e., difficulty in locating a buyer), impediments to the transfer of legal title (viz, the time it takes to verify legal ownership as in a title or patent search and the right to dispose of the asset as in a leasehold interest, dealership, or letter stock), the costs associated with holding the asset, and, most importantly, the price at which you (the owner) are willing to sell. If your minimal price is too dear, then it might never be sold. On the other hand, if the price is exceedingly low (and legal niceties such as proof of ownership are readily established), then the asset might be sold in a very short period of time.

*Department of Economics, University of California, Los Angeles, Los Angeles, CA 90024. This research was partially supported by National Science Foundation grants SES-8308518 and SES-8308479. We acknowledge the helpful comments of Malcolm Fisher, Ronald Masulis, Richard P. Rumelt, E. Roy Weintraub, and Susan Woodward.

[1]His reference to the time dimension is particularly relevant in regard to a premature sale (see Section II, Part C).

Any thoughtful response to clarify the meaning of liquidity must incorporate the idea that the price demanded be "reasonable." The approach suggested here incorporates this idea as it consists in embedding the sale of the asset in a search environment, discerning a sales policy that maximizes the expected discounted value of the net proceeds associated with the sale, and defining the asset's liquidity to be the expected time until the asset is sold when following the optimal policy.[2]

Clearly the concepts of liquidity and money are intimately connected. As defined here an asset's liquidity is the optimal expected time to transform the asset into money. A distinguishing characteristic of money is its role as a medium of exchange.[3] From this perspective, money is desirable because of the ease with which it can be exchanged for other commodities. If we rank commodities by their liquidity, our definition is equivalent to money being the most liquid asset. An exchange of commodity i for commodity j is accomplished most swiftly by first trading i for money and then trading money for j.[4] The expected time to go from i to money corresponds to our measure of i's liquidity. The expected time to go from i to j measures the liquidity of the (i, j) transaction. The crucial point is that in going from commodity i to money, the individual follows an optimal selling policy, and in going from money to j, the individual pursues an optimal buying policy. The approach taken here is novel in that rational behavior under uncertainty, as exhibited by adherence to optimal stopping rules, is *the* defining characteristic of liquidity. This perspective illuminates both the demand for money,[5] and portfolio analysis.[6]

The environment in which the sale of the asset occurs is presented in Section I; there we define the expected time of sale as our measure of an asset's liquidity. The compatibility of this measure with other notions of liquidity is demonstrated in Section II. In particular, we show that our definition is compatible with Keynes (Part A) and that liquidity increases with (i) the market interest rate (Part B), (ii) the thickness of a market with brisk trading (Theorem 1), and (iii) the predictability of offers (Theorem 2). Furthermore, this last result is consonant with the concept of an efficient market. One expects that any change in a parameter which leads to an increase in liquidity also would lead to a decrease in the discount associated with a quick sale. Theorem 3 (Part E) fulfills this expectation for the search cost c.

In the third and final section we proffer a simple search model with a future golden investment opportunity. Theorem 5 reveals that the choice of a more liquid initial investment enhances the investor's ability to profit from the arrival of the golden opportunity: as per our definition, liquidity provides flexibility.

[2] Hirshleifer was the first author to explicitly note the importance of uncertainty and search in determining an asset's liquidity. He forcefully observed: "It is immediately evident that uncertainty is of the essence here"; and "limitations of information may prevent buyers and sellers from finding one another, at least without incurring the costs and uncertainties of a search process" (1968, pp. 1–2).

[3] See Boulding (pp. 310–11) for a lucid discussion of liquidity and the role of money. A deep and influential analysis of the role of money in economic theory is contained in the work of Robert Clower (1977).

[4] Armen Alchian (1977) suggests that the transactions costs in trading i for j will be minimized via trading i for money and money for j with the first trade being effected by a specialist in commodity i and the second trade by a specialist in commodity j. An expanded discussion of this point that includes the importance of search and information is presented in Karl Brunner and Allan Meltzer (1971).

[5] Milton Friedman's demand for money function (1957) contains a variable u that represents uncertainty, among other things. However, uncertainty is a minor actor in his theory of money. Instead of being the tail, Friedman's u variable is now the dog (see p. 9). In a heuristic formulation, Jack Carr and Michael Darby (1981) have based a short-run money demand function upon the effects of money supply shocks on money holdings given reservation prices (presumably the result of optimal search behavior) set by asset sellers and buyers.

[6] Most portfolio analyses (for example, James Tobin, 1958) assume that the appropriate measure of risk is that associated with immediate sale of the assets held in the portfolio. But immediate sale may not be optimal.

I. The Setting

Search is the fundamental feature of the arena in which the sale of the asset is to take place; the setting is very much akin to the standard job search or house selling model. The search environment is characterized by four objects: c_i, T_i, X_i, and β. First, there are the costs of owning/operating the asset as well as the cost of attempting to sell the asset. In the discrete time framework, the net operating and search cost for period i is denoted c_i.

Second, one offer arrives at each time in the set $\{S_i: i=1,2,\dots\}$ of arrival times. The random arrival times S_i satisfy

$$S_i = \sum_{j=1}^{i} T_j,$$

where the integer-valued random variables $T_i \geq 0$ need not be either independent or identically distributed.

The ith price offered is a nonnegative random variable X_i.[7] In the standard search paradigm (see Section II), the X_i are independent, identically distributed, and independent of $\{T_i\}$. None of these three assumptions is invoked here. As evidenced in equation (1) below, our formulation can be structured so that either it does not permit the seller to accept any offer other than the one most recently tendered so recall of past offers is not allowed, or it does permit the seller to accept any of the tendered offers so recall is allowed.

Finally, all expenditures and receipts are discounted[8] at the rate β so that the present value of a dollar received in period i is β^i. The seller seeks to maximize the expected discounted value of his net receipts. More formally, the discounted net receipts $R(\tau)$ associated with a stopping time τ is given by

$$\text{(1a)} \quad R(\tau) = \beta^{\tau} Y_{N(\tau)} - \sum_{i=1}^{\tau} \beta^i c_i$$

$$\text{(1b)} \quad Y_i = \begin{cases} X_i & \text{if no recall} \\ \max(X_1, \dots, X_i) & \text{if recall allowed,} \end{cases}$$

where $N(\tau) = \max\{n: S_n \leq \tau\}$ is the random number of offers that the seller observes when employing the decision rule τ and the random variable $Y_{N(\tau)}$ is the size of the accepted offer. Consequently, the seller chooses a stopping rule τ^* in the set T of all stopping rules (we do not require $P(\tau < \infty) = 1$) such that[9]

$$\text{(2)} \quad ER(\tau^*) = \max\{ER(\tau): \tau \in \mathrm{T}\}.$$

The manifest value of the asset is $V^* \equiv ER(\tau^*)$, and the length of time it takes to realize the asset's value and to convert the asset into cash is the random variable τ^*. (In making this statement we are implicitly assuming that there is no lag between the time an acceptable offer is made and the time that the seller is paid.) We propose to use $E\tau^*$ as the measure of an asset's liquidity with an increase in $E\tau^*$ corresponding to a decrease in liquidity.

The key point is that for any given asset (with its concomitant cost function c_i, arrival times $\{S_i\}$, and offers $\{X_i\}$), there is an optimal policy τ^* which determines the value $ER(\tau^*)$ of the asset. The asset's liquidity is determined by τ^*.[10]

[7] Though it could be accounted for, the impact of inflation upon asset prices is not considered in this analysis.

[8] If the time to sale is relatively short, then discounting will have little impact—though an individual with special short-run opportunities or critical consumption needs could have a high time value of money.

[9] We unabashedly assume the existence of an optimal rule. The assumptions of the standard search paradigm (see the discussion in Section II) ensure the existence of an optimal rule.

[10] Simplicity led us to elect the mean of τ^* as our measure of liquidity. Another increasing function of the distribution of τ^* could have been selected. In particular, the comparative statics results found in Section II, Parts A–D, remain unchanged by any such selection.

(continued)

II. Compatibility with Other Notions of Liquidity

Our measure of liquidity is not only internally consistent but also compatible with a good deal of what economists have said. In regard to its consistency we note that $E\tau^* = 0$ for money so that money is perfectly liquid: it is the most liquid asset.

Second, an illiquid asset is one that can't be sold, or rather one with $E\tau^* = \infty$. This can occur when there are informational asymmetries or structural constraints that induce the potential buyers to undervalue the asset; that is, its worth to the current owner exceeds its assessed or actual worth to any potential buyer. Informational asymmetries arise in the context of a business in which there are many cash transactions and the company's books are not a reliable guide to revenues. Structural constraints such as tax considerations in which only some assets "may be burdened by transaction duties" (Hirshleifer, 1972, p. 137) provide another example of impaired marketability which can render an asset totally illiquid.[11]

To analyze how this might come about, suppose that c_i, the net search and operating cost per period, is $c < 0$ for all i. In addition, suppose that no buyer is willing to offer more than $-c\beta/(1-\beta)$. The policy τ^* of never accepting an offer at or below $-c\beta/(1-\beta)$ yields the owner an expected discounted value of $-c\sum_{i=1}^{\infty}\beta^i = -c\beta/(1-\beta)$ so τ^* is indeed optimal. Moreover, $\tau^* \equiv \infty$ so $E\tau^* = \infty$, and the asset is illiquid.

The "standard" search paradigm is utilized extensively in the ensuing analysis. It entails (i) a constant search cost so $c_i \equiv c$, (ii) one offer tendered each and every period so $T_i \equiv 1$, (iii) independent offers X_i drawn from the same known probability distribution F, and (iv) recall of past offers. With these assumptions the existence of an optimal rule τ^* is guaranteed, and τ^* has the following representation: $\tau^* = \min\{n: X_n \geq \xi\}$, where ξ is referred to as the reservation price. Thus, the seller accepts the first offer greater than or equal to his reservation price ξ; consequently, τ^* is a geometric random variable with parameter $P(X_1 \geq \xi)$, the probability that a given offer is successful in effecting the asset's sale. Furthermore, it is clear upon reflection and easy to demonstrate that $\xi = V^*$. (See our 1976a, b papers for a full discussion of the standard search model and its variants.)

A. *Compatibility with Keynes*

According to Keynes, one asset is more liquid than another if it is "more certainly realisable at short notice without loss" (1930, p. 67). In the context of the standard search paradigm, Keynes' definition is equivalent to ours if we interpret "at short notice," "more certainly realisable," and "without loss" to mean "in one period," "has a higher probability p of being sold in one period," and "in accord with the optimal policy." To see this, recall that the asset is sold if and only if the offer price is ξ or larger, and merely observe that $p = P(X_i \geq \xi)$ is related to $E\tau^*$ via $E\tau^* = 1/p$ so that liquidity increases with p.

B. *Liquidity and Impatience*

Continuing with the standard search paradigm, recall that the reservation price ξ_β is a function of the discount factor β. As per equation (13) of our earlier paper (1976a), the asset's reservation price satisfies

$$(3) \qquad c = H(\xi) - \xi(1-\beta)/\beta,$$

where $H(x) = \int_x^\infty (y-x)\,dF(y)$. Differentiating the first-order condition (3) with respect to β yields

$$(4) \quad d\xi/d\beta = \xi/\beta^2[1-F(\xi)+(1-\beta)/\beta] > 0.$$

Hence, ξ_β is a strictly increasing function of β. Consequently, an increase in β leads to an

This invariance is due to the first-order stochastic dominance of the time τ^* until the sale of the asset induced by the increase in the reservation price ξ. But any analysis in the vein of Theorem 3 would thereby be rendered much more complicated.

[11] A structural characteristic leading to this situation arises when the asset is a business and the current owner's managerial talents in running this business substantially exceed his talents (and implicit wages) in any other employment.

increase in $E\tau^*$ as $E\tau^* = 1/P(X_1 \geq \xi_\beta)$ with ξ_β strictly increasing in β. That is, an increase in the market interest rate or in the asset holder's time preference (a lower value of β) leads to an increase in the asset's liquidity. This demonstrates two facts. First, because more impatience (as might arise from increased consumption needs that only can be satisfied via the expenditure of wealth in the form of money) leads to a more liquid asset, impatience and liquidity preference are commensurate in that they vary directly. As expected, an increase in liquidity preference leads to an increase in liquidity itself. Second, because an asset's liquidity depends upon the discount factor, it is a property of the asset holder as well as an intrinsic property of the asset itself. Nevertheless, to the extent that dispersion of the offer price distribution, rate of receipt of offers, and relative costs of soliciting offers are common across all sellers, the ranking of assets in terms of liquidity will be similar across individuals, regardless of their degree of impatience. With this view we return to the notion that liquidity is determined by characteristics of the asset; characteristics of the seller have virtually no impact.

C. *Liquidity and Thickness of the Market*

When there are many transactions per day of a homogeneous asset such as wheat or long-term Treasury bonds, the market for the asset is thick. On the other hand, the more idiosyncratic the asset, as is the case if it is one-of-a-kind (a work of art or a castle), or has a limited set of uses (a germ-free, refrigerated warehouse or a special purpose lathe), the thinner the market becomes. The number of transactions in a market is a function of several factors, including the frequency of offers received by any particular asset. Accordingly, the thickness of the market for an asset is said to increase with the frequency of offers. Theorem 1 below establishes the direct, though weaker than anticipated, connection between our measure of liquidity and the thickness of the market.

THEOREM 1: *An increase in the frequency of offers causes the expected time of sale to decrease (and liquidity to increase) if either the interest rate is near zero or the frequency of offers is very high.*

PROOF:

To begin the analysis, suppose offers arrive according to a Poisson process with rate λ, and let α be the continuous-time interest rate. Then β_λ, the one-period (a period is the time interval until the next offer) discount factor, satisfies $\beta_\lambda = \lambda/(\alpha+\lambda)$. Define T_λ and τ_λ^* to be the time of sale and the number of offers received until the sale, respectively, when offers arrive at rate λ. As $1/\lambda$ is the expected time between offers, we have $ET_\lambda = E\tau_\lambda^*/\lambda$.

Differentiating with respect to λ and utilizing (4) and $E\tau_\lambda^* = 1/[1-F(\xi_\lambda)]$ yields

$$\begin{aligned} dET_\lambda/d\lambda &= -\,E\tau_\lambda^*/\lambda^2 + \lambda^{-1}\,dE\tau_\lambda^*/d\lambda \\ &= -\left[\lambda^2(1-F(\xi_\lambda))\right]^{-1} \\ &\quad + f(\xi_\lambda)[1-F(\xi_\lambda)]^{-2}\lambda^{-1} \\ &\qquad \times (d\xi_\lambda/d\beta_\lambda)\cdot(d\beta_\lambda/d\lambda) \\ &= -\left[\lambda^2(1-F(\xi_\lambda))\right]^{-1} \\ &\quad + \alpha\xi_\lambda f(\xi_\lambda)\left\{\lambda^3[1-F(\xi_\lambda)]^2 \right. \\ &\qquad \left. \times [1-F(\xi_\lambda)+\alpha/\lambda]\right\}^{-1}. \end{aligned}$$

Noting that both the reservation price ξ and the expected time $[1-F(\xi)]^{-1}$ until an acceptable offer is received are bounded for α near zero and λ large, the above expression for $dET_\lambda/d\lambda$ reveals that its sign is negative when α is near zero and also when λ is large.

The expected time of sale is the product of the expected time between offers and the expected number of offers received until the asset is sold. A decrease in the expected time $1/\lambda$ between offers, the first term in the product, causes the discount factor β_λ to increase. This leads to an increase in the reservation price and hence to an increase in the number of offers received, the second term in the product. Theorem 1 asserts that the net effect is negative if discounting has

little impact (because the time to sale is short or the interest rate is small). In a practical sense, Theorem 1 implies that liquidity, as defined here, increases with thickness for all of the familiar highly organized markets (such as the NYSE) characterized by brisk trading; Theorem 1 is uninformative as regards thin markets.

D. *Liquidity and Predictability*

In Jacob Marschak's view, the word liquidity "denotes a bundle of two measurable properties and is therefore itself not measurable" (1938, p. 323). The two properties he refers to are "plasticity," that is, the ease "of manoeuvring into and out of various yields after the asset has been acquired," and "the low variability of its price." A version of this view of liquidity might provide the following definition: an asset is liquid if it *can be sold quickly* at a *predictable price*. By this definition, liquidity is a two-dimensional attribute.

Consider commodities such as wheat and long-term Treasury bonds. The market for both assets is nearly perfect in that the attempt to sell even as much as one million dollars worth of these assets will have only a minute effect upon "the market price." Moreover, there is a ready (and highly organized) market for both assets with a multitude of transactions taking place each weekday. The transaction can be effected in a matter of minutes. Consequently, it is indisputable that these assets can be sold quickly. On this dimension they would be seen to be near-money.[12]

Recently, however, interest rates have been highly volatile; fluctuations of as much as 9 percent in a single day (recall Federal Reserve Chairman Volcker's announcement of October 6, 1979) have occurred. And the wheat market has a long history of volatility. Thus, neither of these assets rates high on the dimension of "predictable price."

Predictability, we maintain, is an expression of concern with adverse events or downside-risk, that is, safety. As such it ignores and fails to account for the occurrence of favorable events or upside-risk. Our measure of liquidity implicitly utilizes both the adverse and the favorable events by requiring that the asset be sold at its "fair market price" where the price is derived from the seller's optimization (see equations (1) and (2)).

To see the relation between predictability and our measure of liquidity, let $W_i = X_i - \mu$ so $EW_i = 0$ and parameterize predictability by the following representation of the offers: $X_i = \mu + \varepsilon W_i$.

Naturally, a decrease in ε is interpreted as an increase in predictability. An increase in ε is a mean-preserving increase in risk of the sort that might properly be labeled a dilation. We shall limit our investigation to dilations because other mean-preserving increases in risk are less regular in that the concomitant change in liquidity they induce can be either an increase or a decrease.

The seller's problem is to choose a stopping rule τ_ε in the set T of all stopping rules to maximize

$$\begin{aligned} E[\beta^\tau(\mu + \varepsilon W_\tau) - c(\beta + \dots + \beta^\tau)] &= -(c\beta/(1-\beta)) + E\beta^\tau \\ &\quad \times [\mu + (c\beta/(1-\beta)) + \varepsilon W_\tau] \\ &= -(c\beta/(1-\beta)) + \varepsilon E\beta^\tau \\ &\quad \times [(\mu + (c\beta/(1-\beta)))/\varepsilon + W_\tau]. \end{aligned}$$

Equivalently, the seller seeks to maximize

$$E\beta^\tau(\mu_\varepsilon + W_\tau), \tag{5}$$

where $\mu_\varepsilon = (\mu + (c\beta/(1-\beta)))/\varepsilon$. When there is total predictability, that is, when $\varepsilon = 0$,

[12] One might say that such an asset is perfectly marketable. Not only is the owner capable of effecting a quick sale, there is nothing to be gained (on average) from waiting for a better price; i.e., a quick sale can be effected at the market price. This raises the question of whether the concept we have provided measures liquidity or marketability. In our view, this question is largely semantic.

$\tau_\varepsilon \equiv 1$ if $\mu > -c\beta/(1-\beta)$; otherwise, $\tau_\varepsilon = \infty$. In view of this fact we shall assume that $\mu > -c\beta/(1-\beta)$ so that $\mu_\varepsilon > 0$ and μ_ε decreases as ε increases.

It is our intention to show that an increase in the mean μ_ε induces an earlier sale, and, concomitantly, an increase in liquidity.

The stopping problem expressed in (5) is the discounted version of the standard job search problem with a search cost of zero. When $EX_1 = \mu$, the solution is a reservation price ξ_μ such that the seller accepts the offer if and only if it equals or exceeds ξ_μ. As demonstrated in our earlier paper (1976a, p. 164), ξ_μ is the unique solution to

$$(6) \qquad 0 = H_\mu(x) - rx,$$

where $\beta = 1/(1+r)$ and $H_\mu(x) = H(x-\mu)$.[13]

The decreasing nature of H (recall $H'(x) = -(1-F(x))$) yields the following two facts.

LEMMA 1: *If $\delta^+ > \delta$, then $\xi_{\delta+} > \xi_\delta$.*

PROOF:

From the definition of H_δ we have

$$\begin{aligned} H_{\delta+}(\xi_\delta) - r\xi_\delta &= H(\xi_\delta - \delta^+) - r\xi_\delta \\ &= H(\xi_\delta - \delta) - r\xi_\delta + H(\xi_\delta - \delta^+) \\ &\quad - H(\xi_\delta - \delta) \\ &= H(\xi_\delta - \delta^+) - H(\xi_\delta - \delta) > 0. \end{aligned}$$

LEMMA 2: *If $\delta^+ > \delta$, then $\xi_{\delta+} - \xi_\delta < \delta^+ - \delta$.*

PROOF:

By the definition of H_δ,

$$\begin{aligned} H_{\delta+}(\xi_\delta + \delta^+ - \delta) - r(\xi_\delta + \delta^+ - \delta) \\ = H(\xi_\delta - \delta) - r\xi_\delta - r(\delta^+ - \delta) \\ = -r(\delta^+ - \delta) < 0, \end{aligned}$$

so $\xi_\delta + \delta^+ - \delta > \xi_{\delta+}$.

THEOREM 2: *The asset's liquidity is an increasing function of its predictability; that is, $E\tau_\varepsilon$ is an increasing function of ε.*[14]

PROOF:

Fix $\varepsilon^+ > \varepsilon$ so that $\delta^+ \equiv \mu_{\varepsilon+} < \mu_\varepsilon \equiv \delta$. Applying Lemma 2 (with the roles of δ^+ and δ reversed there), we obtain

$$\begin{aligned} p^+ &\equiv P(W_1 + \delta^+ \geq \xi_{\delta+}) \\ &= P(W_1 \geq \xi_{\delta+} - \delta^+) < P(W_1 \geq \xi_\delta - \delta) \equiv p \end{aligned}$$

so that $E\tau_{\varepsilon+} = 1/p^+ > 1/p = E\tau_\varepsilon$.

As the connection between the predictability of the asset's price and the thickness of the asset's market, though presumably direct, is tenuous, we dispense with further comment on this connection. The connection between Theorem 2 and the notion of efficient markets does merit discussion. Theorem 2 states that liquidity decreases with the asset's risk so, absent further specification of the source of risk, it might appear that Theorem 2 contradicts the concept of an efficient market. In a fully informed market there is nothing to be gained by waiting to sell a risky asset, whereas Theorem 2 implies an optimal waiting time which is strictly positive and strictly increasing with the asset's risk. In our analysis the source of uncertainty emanates from the random selection of the particular agent seeking to purchase the asset; in particular, the agents value the asset differentially. Accordingly, the arrival of a

[13] If $P(W_i \leq y) = F(y)$ and $F_\mu(y) = P(W_i + \mu \leq y) = F(y-\mu)$, then

$$\begin{aligned} H_\mu(x) &= \int_x^\infty (y-x)\, dF_\mu(y) = \int_x^\infty (y-x)\, dF(y-\mu) \\ &= \int_{x-\mu}^\infty (z - (x-\mu))\, dF(z) = H(x-\mu). \end{aligned}$$

[14] The reverse result holds if we assume $\mu < -c\beta/(1-\beta)$.

low offer correctly has no impact upon the owner's opinion (an opinion shared by the market) of the asset's value, for the offer does not constitute new information. On the other hand, the source of variation in an efficient markets setting is the arrival of new information. As new information arrives, say in the form of a low offer, each agent, including the asset owner, simultaneously revalues the asset. In short, risk is the embodiment of heterogeneous preferences in one analysis and the arrival of new, commonly shared information in the other. In view of this discussion, it is clear that financial assets traded in a thick, efficient market will be exceedingly liquid.

E. *Liquidity and the Discount Attending Premature Sale*

For some, liquidity corresponds to the following idea of discount.[15,16] Suppose an asset has a value v, but the likely price at which it can be sold in a "quick sale" is only $(1-d/100)v$. Then the discount d associated with the quick sale measures the asset's liquidity—the higher the value of d, the less liquid the asset.

We can incorporate this idea in our search setting. To do so, interpret a quick sale as a constraint that the conversion to cash takes place within a fixed (and perhaps short) amount of time t.[17] Then only policies in the set $T_t = \{\tau \in \mathrm{T}: \tau \leq t\}$ are permitted. Hence, the seller seeks a stopping rule τ_t in the set T_t such that

$$(7) \quad V_t \equiv ER(\tau_t) = \max\{ER(\tau): \tau \in T_t\}.$$

The corresponding discount is $100(1-V_t/V^*)$ which we label $d(t)$.

Makower and Marschak describe their concept of saleability "as the relationship between the selling price and the time which the seller must wait in order to get it" (p. 280). Continuing in this vein, they state that "the influence of time on the selling price is due to the seller's finding more buyers." With these ideas, their deterministic "price-time schedule" is very much akin to the function V_t and the waiting for offers in the search environment is not very different from their idea of waiting in order to find more buyers.

This formalization of liquidity is not necessarily the same as the one proposed earlier, for it can easily happen that the discount $d_1(t)$ for asset 1 is less than the discount $d_2(t)$ for asset 2, yet $E\tau_1^* > E\tau_2^*$. More generally, $d_1(t)-d_2(t)$ can change sign as t increases.

While we readily acknowledge that there are instances in which our proposed measure $E\tau^*$ of liquidity is not commensurate with the notion of liquidity embedded in the discount $100(1-V_t/V^*)$, we expect that these two measures will agree frequently. In fact, just as an increase in the cost c of search leads to an increase in liquidity as measured by the expected time to sale, we demonstrate in Theorem 3 that an increase in c also causes the discount $100(1-V_n/V^*)$ to decrease for all horizon lengths n. Even though these two measures are not mathematically equivalent, this result suggests they are compatible in a practical sense.

THEOREM 3: *In the context of the standard search paradigm with recall, an increase in the*

[15] Hirshleifer asserts that "Illiquid assets...are those characterized by a relatively large discount for 'premature' realization" (1972, p. 137).

[16] Roland McKean uses liquidity "to mean merely 'moneyness'," and asserts that "Usually, an asset's liquidity is described to include the probabilities of getting various fractions of the going price plus the time period necessary to liquidate the asset" (1949, p. 509). These "fractions" correspond to our idea of discount. Like Marschak, McKean believes that an operational definition of liquidity is not possible:

> Since these components cannot be measured, there is little to be gained by breaking the notion down. Perhaps it is sufficient to say that the more nearly we regard an asset as substitutable for money, or the more it partakes of the same attractions possessed by money-holdings, the more liquidity the asset has. [p. 509–10]

[17] Alternatively, suppose that the optimal stopping rule is preempted by an event requiring immediate disposal of the asset. This event can be interpreted as either a once-in-a-lifetime investment opportunity or a catastrophe. The time at which preemption occurs is a random variable and can be included in the formulation of the stopping rule problem.

cost of search causes both $E\tau^*$ and $(1-V_n/V^*)$ to decrease, $n=1,2,\ldots$, where V_n, defined in (7), is the value of the asset when it must be sold within n periods.

PROOF:

Differentiating the first-order condition (3) with respect to c yields

$$\xi' = -\beta/[1-\beta F(\xi)] < 0. \tag{8}$$

Hence, an increase in c causes ξ, and in turn $E\tau^* = 1/P(X_1 \geq \xi)$, to decrease.

It is clear upon reflection that ξ represents not only the asset's reservation price but also its value; that is, $\xi = V^*$. Consequently, in order to demonstrate that the discount decreases with c it suffices to show

$$d[V_n/\xi]/dc \geq 0, \qquad n=1,2,\ldots. \tag{9}$$

To begin the analysis it behooves us to notice that

$$V_1 = \beta(\mu - c) \tag{10a}$$

$$V_{n+1} = -\beta c + \beta F(\xi) V_n + \beta \int_\xi^\infty x\, dF(x), \qquad n \geq 1 \tag{10b}$$

where $\mu = EX_1$ and we have used the fact (see our 1976a paper, p. 170) that the reservation price when n periods remain is ξ, $n=1,2,\ldots$. (This fact provides an enormous simplification in the analysis vis-à-vis the case of no recall.) From (8) and (10a) we obtain

$$(d/dc)[V_1/\xi] = \left\{-\xi + \frac{\beta(\mu-c)}{1-\beta F(\xi)}\right\}\beta/\xi^2, \tag{11}$$

whereas manipulation of (3) produces

$$\mu - c = (\xi/\beta) - \int_0^\xi (\xi - x)\, dF(x). \tag{12}$$

Inserting (12) into (11) generates

$$\xi^2(1-\beta F(\xi))(d/dc)[V_1/\xi] = \beta^2 \int_0^\xi x\, dF(x). \tag{13}$$

The nonnegativity of $d[V_1/\xi]/dc$ is palpable from its representation in (13).

To simplify the rather complex expressions in $d[V_n/\xi]/dc$, we shall write F and f in place of $F(\xi)$ and $f(\xi)$ and $D \equiv \xi F(\xi) - \int_0^\xi x\, dF(x)$. From (10) and (12) we have

$$\begin{aligned}(V_2 - V_1)/\beta &= \beta F(\mu - c) - \mu + \int_\xi^\infty x\, dF(x) \\ &= \beta F\left[\xi/\beta - \int_0^\xi (\xi - x)\, dF(x)\right] - \int_0^\xi x\, DF(x) \\ &= (1-\beta F)D,\end{aligned} \tag{14}$$

whereas iterating the first differences obtained via (10) leads to

$$\begin{aligned}V_{n+1} - V_n &= \beta F(V_n - V_{n-1}) = \ldots \\ &= (\beta F)^{n-1}(V_2 - V_1), \quad n \geq 1.\end{aligned} \tag{15}$$

From (14) we easily realize

$$(d/dc)(V_2 - V_1) = \beta F(1-\beta F) + \beta D - \beta^2 D f \xi'. \tag{16}$$

Employing (8), (14), and (16) in conjunction with (15) yields

$$\begin{aligned}&\xi^2(d/dc)[(V_{n+1} - V_n)/\xi] \\ &= (\beta F)^{n-1}\{\xi\beta F(1-\beta F) + \xi\beta D + \beta^2 D\} \\ &\quad + \beta^2 D \xi (\beta F)^{-1} f \xi' \\ &\times \{(\beta F)^{n-1}[n(1-\beta F) - 1]\}, \quad n = 1,2,\ldots\end{aligned} \tag{17}$$

As differentiation is a linear operator and $V_{n+1} = \Sigma_{i=1}^n (V_{i+1} - V_i) + V_1$, equations (16)

and (17) enable us to conclude that $[\gamma \equiv \beta F]$

$$(18)\quad \xi^2(1-\beta F)(d/dc)[V_{n+1}/\xi] = (1-\gamma^n)[\xi\beta F(1-\beta F)+\xi\beta D+\beta^2 D] + \beta^2\int_0^{\xi} x\,dF(x) - \beta^3 D\xi(\beta F)^{-1} \times f\left\{(1-\gamma)\sum_{i=1}^{n} i\gamma^{i-1} - \sum_{i=1}^{n}\gamma^{i-1}\right\}.$$

Because the term in braces equals $-n\gamma^n < 0$ and $D \geq 0$, all of the terms on the right-hand side of (17) are nonnegative.

The value of Theorem 3 resides in its demonstration of the compatibility of the two measures rather than in the conclusion that an increase in the cost of search leads to an increase in liquidity.[18] In fact, this conclusion is somewhat counterintuitive. We offer two distinct arguments to diminish the disturbing aspects of this counterintuitive result.

First, there need be no connection between costly offers and infrequent offers. For instance, if the asset earns a large net rent and there is an out-of-pocket expense associated with obtaining an offer, then an increase in the frequency of offers could change the sign of the search cost from negative to positive.

Second, when properly viewed, this result raises nary an eyebrow in a labor market context. Theorem 3 asserts that the expected duration of unemployment is shorter for workers with high search costs: their reservation wage is lower; hence they more readily accept offers of employment. If worker B has a higher search cost than A, his (expected) period of unemployment is shorter. If A can signal his desirability more easily than B, then we anticipate that A will have a lower search cost and, therefore, a longer duration of unemployment. In the same vein, suppose C has the same search cost as A but C is a less able worker. In particular, suppose each offer received by C is δ less than the corresponding offer received by A. As shown in Theorem 4, the expected duration of unemployment is shorter for A.[19] Clearly, a long period of unemployment is not synonymous with an inferior employee: workers with short periods of unemployment may be the ones with good job prospects (as per Theorem 4) or impaired ability to signal their worth (as per Theorem 3). Similarly, unless all other aspects are identical, the less liquid asset need not be inferior.

THEOREM 4: *Let ξ and ξ_δ denote the reservation wage in the standard search paradigm when the offer distributions F and F_δ satisfy $F_\delta(t) = P(X+\delta \leq t) = P(X \leq t-\delta) = F(t-\delta)$ with $\delta > 0$; thus, each offer in the second problem is δ larger than each corresponding offer in the original problem. If $\beta < 1$, then $\xi_\delta < \xi + \delta$ and τ^* is (stochastically) larger than τ_δ^*. If $\beta = 1$, then $\xi_\delta = \xi + \delta$ and τ^* has the same distribution as τ_δ^*.*

PROOF:

Footnote 13 reveals that $H_\delta(t) = H(t-\delta)$. Suppose $\beta < 1$ and $\xi_\delta \geq \xi + \delta$. Substituting into the first-order condition (3) yields

$$H(\xi) - \frac{1-\beta}{\beta}\xi = c = H(\xi_\delta - \delta) - ((1-\beta)/\beta)\xi_\delta < H(\xi) - ((1-\beta)/\beta)\xi,$$

[18] To ameliorate this counterintuitive result we might use the expected discounted cost of search in place of the expected time of sale as our operational measure of liquidity. However, as shown in ch. 3 of our forthcoming book, this operational measure also can increase with increases in c.

[19] The explanation of this phenomenon is implicit in the proof of Theorem 4: an upward shift in the mean of the offer distribution causes the opportunity cost of search to increase. On the other hand, if F_δ, the offer distribution for A, arises from a multiplicative shift with $\delta > 1$ (so $F_\delta(t) = F(t/\delta)$), then the impact of this shift is easily seen to be equivalent to a decrease in the search cost which induces a longer duration of unemployment. Thus, the direction of change in the duration of unemployment due to an increase in the worker's ability, reflected in his offer distribution, is sensitive to the form in which this increased ability is embodied.

as $H(\cdot)$ is a decreasing function and $\xi_\delta \geq \xi + \delta > \xi$ by assumption. This contradiction reveals that $\xi_\delta < \xi + \delta$. Consequently, $P(X_\delta \geq \xi_\delta) = P(X + \delta \geq \xi_\delta) = P(X \geq \xi_\delta - \delta) \geq P(X \geq \xi)$ so that τ^* is stochastically larger than τ_δ^*.

III. Liquidity as Flexibility[20]

An investment of funds today obviously reduces the range of options open to the agent tomorrow. This flexibility aspect of liquidity is implicit in Section I. Both the search paradigm, in general, and the liquidity search model, in particular, are based on the opportunity cost doctrine—the cost of holding one asset is the return that could be achieved by investing in the next best asset. Most of search theory employs models in which these opportunity costs are constant. However, if the agent's future opportunities differ from his current opportunities, he may eschew commitments that yield an inflexible or illiquid portfolio. In terms of our definition, he may avoid investments with large values of $E\tau^*$.

The key feature of the simple search model we propose is the existence of a single golden investment opportunity that becomes available at some future date. This feature of the investment environment causes the constant opportunity costs to vanish. By specifying the functional form of the offer distribution F, we are able to demonstrate the investor's preference for a more liquid/flexible current investment.

At time 0 the investor's endowment is ξ, all in the form of cash. A set of assets parameterized by $\lambda > 0$ is available for purchase at cost ξ, where asset λ has the associated offer distribution F_λ and search cost c_λ. For simplicity in presentation, assume that each asset's reservation price ξ_λ satisfies $c_\lambda = H_\lambda(\xi_\lambda)$ rather than (3), the discounted form of the first-order condition. The value ξ_λ of asset λ equals its purchase cost ξ, and each asset generates a constant flow of income at the rate $\alpha\xi$, where $\alpha > 0$ is the continuous-time interest rate (i.e., $e^{-\alpha}$ plays the role of β). Assume that the investor purchases exactly one asset at time 0 and consumes its income stream as it flows in. Thus, after selecting an asset, say λ_0, for purchase at time 0, the investor's endowment is λ_0 at each point in time.

At *one* point T in time in the future, the investor will be presented with the opportunity to invest in the golden asset. He will have enough time to solicit exactly one offer for the asset he owns in order to generate cash to purchase the golden asset. For the purposes of our analysis it does not matter if T is a random variable with known distribution, deterministic and specified in advance, or uncertain in the sense of Frank Knight (1921). What is important is that there be time to generate but one offer. (This will be the case if T is a geometric random variable.)

The golden asset is divisible and has constant returns to scale. Each dollar invested in the golden asset generates a constant flow of income at the rate r per unit time. As implied by its name, the income flow associated with an investment of ξ dollars in the golden asset exceeds $\alpha\xi$; that is, $r > \alpha$.

Let $G(\lambda)$ be the expected gain to search at time T when asset λ was purchased at time 0. If the observed value x of the offer X_λ were to be invested in the golden asset, the resulting cash flow would be rx. This investment is made only if $rx > \alpha\xi$; otherwise, the investor foregoes investment in the golden asset and retains asset λ. Hence, $\max\{rX_\lambda/\alpha; \xi\} - \xi_\lambda$ is the return to search, and $G(\lambda)$, the expected gain, is given by

$$
\begin{aligned}
(19)\quad G(\lambda) &= E\max\{rX_\lambda/\alpha; \xi\} - c_\lambda - \xi \\
&= \frac{r}{\alpha}E\max\{X_\lambda; \alpha\xi/r\} - c_\lambda - \xi \\
&= \frac{r}{\alpha}\left\{\frac{\alpha\xi}{r}F_\lambda\left(\frac{\alpha\xi}{r}\right) + \int_{\alpha\xi/r}^{\infty} x\,dF_\lambda(x)\right\} - c_\lambda - \xi \\
&= \frac{r}{\alpha}\left\{\frac{\alpha\xi}{r} + \int_{\alpha\xi/r}^{\infty}\left(x - \frac{\alpha\xi}{r}\right)dF_\lambda(x)\right\} - c_\lambda - \xi \\
&= (r/\alpha)H_\lambda(\alpha\xi/r) - c_\lambda,
\end{aligned}
$$

[20]The discussion in this section is in the spirit of Albert Hart (1942), John Hicks (1974), and, especially, Robert Jones and Joseph Ostroy (1984).

where H_λ is the usual H function associated with the offer distribution F_λ.

Recalling that $r > \alpha$ and H_λ is a nonincreasing function, we observe that

$$G(\lambda) > H_\lambda(\alpha\xi/r) - c_\lambda \geq H_\lambda(\xi) - c_\lambda = 0,$$

so search is profitable for each asset λ.

Most of us believe that liquidity is sought to provide flexibility—be it to meet special consumption exigencies or special (golden) investment opportunities. In order to test the validity of this conventional wisdom in the context of our simple search model with a golden opportunity, we shall assume further that the offer distributions F_λ are exponential: $F_\lambda(x) = 1 - e^{-\lambda x}$. Consequently, $H_\lambda(x) = e^{-\lambda x}/\lambda$ and $c_\lambda = e^{-\lambda\xi}/\lambda$. With exponential offers we have

$$E\tau_\lambda^* = [1 - F_\lambda(\xi_\lambda)]^{-1} = 1/e^{-\lambda\xi}$$

so $E\tau_\lambda^*$ increases with λ: liquidity decreases as λ increases. From (19) and F_λ exponential we obtain

$$G(\lambda) = \frac{r}{\alpha}\frac{1}{\lambda}e^{-\lambda\xi\alpha/r} - e^{-\lambda\xi}/\lambda$$

and then, because $r/\alpha > 1$,

$$\lambda^2\frac{dG(\lambda)}{d\lambda} = \frac{r}{\alpha}\left\{-\frac{\lambda\xi\alpha}{r}e^{-\lambda\xi\alpha/r} - e^{-\lambda\xi\alpha/r}\right\}$$
$$-\{-\lambda\xi e^{-\lambda\xi} - e^{-\lambda\xi}\} = (\lambda\xi + 1)e^{-\lambda\xi}$$
$$-(\lambda\xi + (r/\alpha))e^{-\lambda\xi\alpha/r} < 0.$$

Thus, the expected gain to search also decreases with λ. More formally, we have established the following theorem.

THEOREM 5: *In the simple search model (with timeless search), a set of initially available assets with exponential offer distributions, and a subsequently available golden asset, the risk-neutral investor improves his expected return by selecting a more liquid asset for his initial investment.*

As conjectured, the choice of a more liquid initial investment does indeed enhance and facilitate the investor's ability to profit from the arrival of the golden opportunity. We view this result as providing an endorsement of John Hicks' remark that "by holding the imperfectly liquid asset the holder has narrowed the band of opportunities which may be open to him..." (p. 43–44). By choosing a less liquid asset, the investor has more nearly "locked himself in." In particular, note that the probability $(1 - \exp\{-\lambda\xi\alpha/r\})$ of not investing in the golden asset—being locked in—increases as the asset's liquidity decreases.

Although ours is not an equilibrium analysis and the extent to which this result is robust remains to be investigated, our analysis of the simple search model with a golden asset is tantalizingly suggestive of a broad range of macroeconomic phenomena that might successfully be treated with our approach to liquidity and liquidity preference. Though it may be somewhat grandiose, we conclude by paraphrasing Robert Jones and Joseph Ostroy's (p. 26) remarks concerning the profession's long but spotty treatment of flexibility: the difficulty of providing a definition of liquidity in such a way as to have universal application and the difficulty of obtaining formal results without model-specific qualifications may account for the very limited role accorded to liquidity in contemporary theory. However, the connection between liquidity and (risky) investment decisions is too compelling to be ignored.

REFERENCES

Alchian, Armen A., "Why Money?," *Journal of Money, Credit, and Banking*, February 1977, *9*, 133–40.

Boulding, Kenneth E., *Economic Analysis*, 3rd ed., New York: Harper and Brothers, 1955.

Brunner, Karl and Meltzer, Allan D., "The Uses of Money: Money in the Theory of an Exchange Economy," *American Economic Review*, December 1971, *61*, 784–805.

Carr, Jack and Michael R. Darby, "The Role of Money Supply Shocks in the Short-Run Demand for Money," *Journal of Monetary*

Economics, September 1981, *8*, 183–99.

Clower, Robert W., "The Anatomy of Monetary Theory," *American Economic Review Proceedings*, May 1977, *67*, 206–12.

Friedman, Milton, "The Quantity Theory of Money—A Restatement," in his *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1957.

Hart, Albert G., "Risk, Uncertainty, and the Unprofitability of Compounding Probabilities," in O. Lange, F. McIntyre, and T. Yntema, eds., *Studies in Mathematical Economics and Econometrics*, Chicago: University of Chicago Press, 1942.

Hicks, John R., *The Crisis in Keynesian Economics*, New York: Basic Books, 1974.

Hirshleifer, Jack, "Liquidity, Uncertainty, and the Accumulation of Assets," CORE Discussion Paper No. 6810, June 1968.

______, "Liquidity, Uncertainty, and the Accumulation of Information," in C. F. Carter and J. L. Ford, eds., *Essays in Honor of G. L. S. Shackle*, Oxford: Basil Blackwell, 1972.

Jones, Robert A. and Ostroy, Joseph M., "Flexibility and Uncertainty," *Review of Economic Studies*, January 1984, *51*, 13–32.

Keynes, John Maynard, *A Treatise on Money*, Vol. 2, London: 1930.

______, *The General Theory of Employment, Interest and Money*, New York: Harcourt, Brace and Company, 1936.

Knight, Frank, *Risk, Uncertainty and Profit*, New York: Houghton Mifflin, 1921.

Lippman, Steven A. and McCall, John J., (1976a) "The Economics of Job Search: A Survey: Part I," *Economic Inquiry*, June 1976, *14*, 155–89.

______ **and** ______, (1976b) "The Economics of Job Search: A Survey: Part II," *Economic Inquiry*, September 1976, *14*, 347–68.

______ **and** ______, *The Economics of Job Search*, forthcoming.

McKean, Roland N., "Liquidity and a National Balance Sheet," *Journal of Political Economy*, December 1949, *57*, 506–22.

Makower, Helen, and Marschak, Jacob, "Assets, Prices and Monetary Theory," *Economica*, August 1938, *5*, 261–87.

Marschak, Jacob, "Money and the Theory of Assets," *Econometrica*, October 1938, *6*, 311–25.

Tobin, James, "Liquidity Preference as Behavior Toward Risk," *Review of Economic Studies*, February 1958, *25*, 65–86.

Illegal Immigration: The Host-Country Problem

By WILFRED J. ETHIER*

What can the pure theory of international trade tell us about international migration? I have long found it curious that, although migration is a central feature of the international economy and has been for a long time, and although regional economists have devoted considerable attention to the phenomenon, it has never received from the pure theory of international trade more than a small fraction of the attention lavished on the theory of international capital movements. No doubt part of the reason for this is the widely appreciated fact that much of the latter is in effect a theory of factor movements, if for no better reason than that nothing which might distinguish one factor from another is allowed to play a significant role. Thus, one might argue, we do in fact have a significant literature on migration: all that is necessary is that the word "capital" be replaced by "labor" in much of the theory of international capital movements. Indeed, a good part of what trade theorists have written about migration has in effect done just that, with labor distinguished in no essential way from other factors. This approach to migration is all to the good, as far as it goes. It does, after all, give us a useful theory, and it gives us that theory on the cheap. But we are quite noticeably without a theory addressed in any substantive way to those aspects of migration that are distinctive to it.[1]

A notable exception is the recognition that labor movements involve migration of the owner of a factor service as well as of the service itself. This has both positive and normative implications, each addressed in the literature. On the positive side, an export of labor does not generate in exchange a flow of payments, unless wages are repatriated. The analysis to follow in this paper is not sensitive to such an assumption, so I shall not further allude to the issue. The normative implication is that one must decide whether to treat the welfare of migrants as of policy concern to the country from which they have come, to that to which they have gone, or to neither.[2] The host-country perspective of this paper precludes adoption of the first alternative and the illegal status of migration limits the relevance of the second, though it will occasionally become pertinent to what follows.

There are, I believe, three key parameters that in fact distinguish incidents of migration from each other and with which any serious theory of migration must come to grips.[3] The first parameter involves whether the migration is intended to be temporary or permanent. This consideration is central to the role of migrants in both the source and host economies. Although temporary migration seems to be significantly more common in

*Department of Economics, University of Pennsylvania, Philadelphia, PA 19104. This paper has benefited from discussions during seminars at Indiana University, the Institute for International Economic Studies in Stockholm, the International Economics Study Group in London, and Pennsylvania State University, and also from both an International Trade Workshop held at the University of Western Ontario in March of 1984 and a Migration Workshop conducted in March 1985 by the Royal Institute of International Affairs and the Centre for Economic Policy Research in London.

[1] In simple factor-endowments models, international capital mobility is a perfect substitute for international labor mobility; this could limit interest in the explicit consideration of the latter. But, as a referee has pointed out, there are in fact many industries which have no feasible substitute for labor migration: some, like agriculture, use a factor that is not internationally mobile, and others, such as services, produce nontraded outputs.

[2] See, for example, Jagdish Bhagwati (1979).

[3] For a more detailed discussion of the implications of conventional trade theory for migration and of the distinguishing characteristics of the latter, see my paper (1984). The three parameters discussed below are not necessarily of significance *only* for labor mobility. As a referee has noted, the temporary or permanent nature of a flow is often important for the implications of capital movements.

the contemporary world, both types are clearly important in practice.

The second key consideration is whether migration involves skilled or unskilled labor. The former is in effect a simultaneous movement of labor and human capital. The migration of unskilled labor seems more common at present, but, again, both types are important.

The third central parameter is the legality of the migration. Though restrictions might in principle be imposed by both source and host countries, it is the (potential) hosts which in fact do so.[4] An appraisal of the relative quantitative importance of illegal immigration is, by its very nature, hard to get a handle on, but for some countries—notably the United States—it seems to be at least as large as legal migration.

The various permutations of these characteristics are not, of course, of equal importance. Some can be ignored altogether: the illegal migration of skilled workers, both temporary and permanent, is of little interest simply because most countries are quite willing to admit such migrants. The permanent legal migration of unskilled labor is the natural area of application of our existing theories (those, that is, which ignore all distinctive features of labor), although this particular type seems to be of relatively modest practical importance. The migration of skilled labor has spawned its own distinctive literature (that devoted to the "brain drain"). The temporary migration of unskilled labor, both legal and illegal, constitutes the larger part of actual migration. I have elsewhere (1985) investigated the temporary aspect of this. This paper accordingly addresses the notable outstanding gap: the illegal migration of unskilled labor.[5] As the paper ventures into untrodden terrain it will be preoccupied with the formulation of theory. I shall have to make many arbitrary decisions about what to include and what to exclude. One decision is reflected in my title: this paper will confine itself to a host-country perspective.[6]

This focus is primarily a division of labor, not a judgement of what is significant and what is not. As will become apparent, my development of the topic has been strongly influenced by policy issues in the United States. But I submit that this is increasingly convenient from a broader Atlantic perspective as well: the important European problems in this area now look (compared to the 1950's and 1960's, at any rate) more and more like the outstanding American problems, and policy reforms under discussion in the United States sound more and more like European practice.

I. Border Enforcement

The subject of illegal immigration is presumably defined by partially successful attempts to prevent that migration. The modeling of these enforcement efforts should accordingly occupy center stage in the theory to be developed.[7] This section seeks to present, in as simple a context as possible, the salient aspects of border enforcement policy.

A. *A Simple Model of Border Enforcement*

Suppose that skilled workers (S) and unskilled workers (U) can be used to pro-

[4] Communist countries constitute exceptions to this, but very few of our theories have been formulated with regard to their ability to accommodate these states.

[5] Although we do not have a theory of illegal immigration, there is a substantial amount of work on smuggling and illegal trade. See Bhagwati and Bent Hansen (1973), Mark Pitt (1981), and other references cited in the useful overview by Bhagwati (1981). This paper differs from that literature in important ways, beyond the obvious one that I examine factor movements rather than commodity trade. The smuggling literature is concerned with the efforts of importers (sometimes involving the expenditure of real resources) to expand trade by evading tariffs; the enforcement effort (and its real cost) is typically taken as given. I am concerned with the efforts of the host country (involving the expenditure of real resources) to limit the inflow of migrants, and I accordingly treat the degrees of various types of enforcement as central instrumental variables. Perhaps each literature could profit by exploiting the insights of the other, but that is not the purpose of this paper.

[6] For a look at international interdependence, see Slobodan Djajic (1985).

[7] My efforts will be very much in the spirit of the economic theory of crime pioneered by Gary Becker (1968). See also Isaac Ehrlich (1974).

duce output (Q) via a neoclassical production function:

$$Q = Sf(u)$$

where u denotes employment of unskilled workers per skilled worker. Skilled workers are all natives. Unskilled labor consists of two parts, $U = L + I$, with L denoting the supply of native unskilled workers plus legal immigrants, and I the number of illegal immigrants present in the host country.

Of those (M) who illegally attempt to immigrate, a certain number (C) are caught and denied entry,[8] so that $I = M - C$. The authorities' success in preventing illegal entry presumably depends upon the resources devoted to border enforcement.[9] Denote by E the quantity, in terms of output, of such resources. An increase in E would increase the authorities' apprehension ratio, C/M, although probably at a decreasing rate. I accordingly assume that

$$C/M = g(E)$$

where $g(0) = 0, g' > 0, g'' < 0, g < 1$.

Although I have announced my intention to stick to a host-country perspective, it is crucial that the decision to migrate be explicitly modeled in some way. I assume that foreign workers in the source country have the choice of remaining there and earning the wage (or wage equivalent) w^*, which I take to be exogenous, or of attempting to migrate. If successful they earn the expected wage $\tilde{w}$, received by host-country unskilled workers, because either legal or illegal migrants are indistinguishable to firms or else the distinction is of no consequence to them.[10] If unsuccessful, the would-be migrants earn $w^* - k$, where k is the (wage equivalent of) the penalty suffered by those who are caught. The probability of a typical potential migrant getting caught is C/M, or $g(E)$. Assuming that foreign workers are risk neutral, attempted migration adjusts so as to equate the expected reward from migration to the local wage:

$$(w^* - k)g + \tilde{w}(1-g) = w^*.$$

Thus we have

$$(1)\quad \tilde{w} = w^* + k[g(E)/(1-g(E))] = w(E)$$

It follows then that the expected wage of unskilled labor is directly determined by the effort put into border enforcement. The submarkets in which illegal migrants are important are usually characterized by relatively high unemployment rates. I suppose therefore that the unskilled labor market has a rigid wage, w, presumably at a level above the market-clearing one. Suppose that this market operates in a familiar Harris-Todaro fashion:[11] firms employ unskilled labor to the point at which the rigid wage just equals the value of the worker's marginal product, and the consequently scarce jobs are allocated among workers by means of a random draw, so the expected wage $\tilde{w}$ confronted *ex ante* by unskilled workers is ew, where $e(I+L)$ denotes the total employment of unskilled labor. Then

$$(2)\qquad w = f'(u),$$

$$(3)\qquad ew = w(E),$$

where $u = e(I+L)/S$ and $w(E)$ is as in (1). The market for skilled workers, by con-

[8] The variable M actually denotes the total number of attempts to migrate, not the number of individuals making such attempts. Some of those who at first do not succeed will presumably try again.

[9] I have in mind illegal entry, rather than illegal "overstaying." While the former is the main focus of interest in the United States, the latter is relatively more important in some European countries. Much of my analysis does in fact apply to overstaying as well, but I will not explicitly address that form of illegal immigration.

[10] This point will be investigated in great detail later in the paper.

[11] See John Harris and Michael Todaro (1970). A similar assumption has been made when unemployment has been considered in the brain drain literature: see Bhagwati and Koichi Hamada (1974) and Carlos Rodriguez (1975). See also Djajic.

trast, is assumed to clear, and they are paid a wage v reflecting the value of their marginal product:

$$(4) \qquad v = f(u) - uf'(u).$$

I assume that E is financed by a tax on the wages of skilled workers only:

$$(5) \qquad E = tvS.$$

Now E determines e from (1) and (3), w determines u from (2), (4) gives v, and (5) determines the tax rate t. Then each value of E yields a unique equilibrium, and an increase in E will raise e, reduce I, require a larger t, and leave u and v unchanged. Such a policy change consequently harms skilled workers and benefits employed unskilled workers, both those in the country legally and those there illegally. The improvement takes the form of better job prospects, in view of the assumed rigid wage. Potential migrants are neither better off nor worse off in an *ex ante* sense, although the smaller number who actually succeed in entering the country do then have improved prospects.

Note that equations (1), (2), and (3) imply that

$$\tilde{w} - w^* = [uwS/(I+L)] - w^*.$$

Thus increases in the wage gap, $\tilde{w} - w^*$, will be associated with *decreases* in the actual volume I of illegal immigration. A large part of the substantial empirical literature on migration proceeds, by contrast, from the presumption that actual migration ought to be positively correlated with the wage gap.

B. *Objectives of Enforcement*

Why should a nation have an interdiction policy? The international capital movements literature, with its emphasis on the optimal rate of taxation of foreign earnings, suggests one answer: to exploit market power with respect to mobile labor; that is, to maximize national income. A second policy objective could be the internal distribution of income between skilled and unskilled workers. A final objective consistent with the present model is I itself; the authorities might wish to control the number of immigrants in the country for social reasons.

What policy instruments are available to the authorities to help achieve these three targets? There is, of course, border interdiction, E. The tax rate t is clearly not a distinct policy tool. However, L, the number of legal unskilled workers, is, because it is determined by the authorities' decision about how many migrants to admit legally. Thus, in the context of the present model, there are apparently three potential policy targets and two potential instruments.

The actual situation, however, differs in several ways from the potential one. Consider the first target, the maximization of national income. One would expect border enforcement to be inferior to a tax on migrant wages, since the former uses up revenues instead of generating them.[12] But a migrant wage tax may not be feasible—much is infeasible in this area of inquiry. How then, would an increase in the border interdiction effort E affect host-country national income? To see this, measure home real income as

$$(6) \qquad Y = weL + vS - E.$$

Note that I include the wages of native unskilled workers and legal migrants, but exclude wages paid to illegal immigrants. Now an increase in E will have no effect on w, L, v, or S, but will raise e (benefiting unskilled workers) and E (harming skilled workers). National income will obviously rise if and only if

$$[weL/E]\,\hat{e} > \hat{E},$$

where a circumflex denotes proportional change (for example, $\hat{e} = de/e$). Differentiation of (1) and (3) reveals that

$$\hat{e} = [(ew - w^*)/ew]\,\varepsilon_B/[1 - g(E)]\,\hat{E},$$

where ε_B denotes the elasticity of the border

[12] A fee for temporary migration has been advocated by Edwin Harwood, in Nathan Glazer (1985).

enforcement schedule (Eg'/g). Thus an increase in the interdiction effort E will succeed in raising host-country national income only if an increase in E is sufficiently effective so that

$$(7) \qquad \varepsilon_B > [1-g]E/[we-w^*]L.$$

If immigration policy does in fact raise national income, it will not do so by exploiting monopoly power in the international labor market: I have assumed away such power, and migrant workers, like native unskilled workers, are paid the rigid wage w in any event. Instead, the possibility of an increase in national income reflects the assumed distortion in the host unskilled labor market; if migrants are successfully excluded, there are more jobs available at an unchanging wage for legal workers. Indeed the possession of monopoly power in the world market for unskilled workers could easily make it *more* difficult for an increase in E to raise national income! For suppose that a rise in E, by making attempts at migration less tempting to foreign workers and therefore inducing more to remain in the source country labor market, lowers w^*. This in turn makes staying at home less attractive: the rise in E is less effective in reducing the temptation to migrate when w^* can fall than when it is unresponsive. Thus fewer jobs could be freed for legal residents,[13] and the condition for host-country national income to rise would accordingly become stricter than (7).

This analysis should leave us, I believe, with considerable skepticism about the ability of interdiction policy to influence national income. With any effect on the international market for unskilled workers likely to be counterproductive, the full burden falls on ε_B. But skepticism about the ability of marginal changes in E to have significant effect is widespread in practice. Furthermore, in this model national income can be raised only by redistributing income from skilled to unskilled workers. With this distribution itself a policy target there is no freedom to influence national income even if it is feasible to do so. Finally, it is the presence of a distorted home labor market that makes it possible to use E to raise Y. One would expect, therefore, that this possibility would disappear if instead w were to adjust to clear the market for unskilled labor.[14]

With immigration policy ineffective in the exploitation of market power in the migrant labor market, we would appear to be left with two targets and two instruments. But this is not so. Although L can be varied directly, it is not, in fact, a meaningful instrument. As we have seen, the rigid wage w directly determines the technique of production, $u = e[L+I]/S$, with e determined by E. It follows, therefore, that any change in L will simply cause an offsetting change in I, and thus in C and M also. Reducing the number of legal immigrants will increase by a greater amount those who attempt illegal entry, and by an equal amount those who succeed. We have, therefore, only one policy instrument.

We still have two potential targets, e (or the internal *ex ante* distribution of income) and I. But note that equation (2) inextricably binds these two targets together: with u set by the rigid wage, e directly determines I. There is no way to unbundle these two targets. Unless the authorities just happen to desire a consistent combination of e and I, they will have to trade off one goal for the other, even before considering what tools might be available.

This, then, is the policy environment implied by my simple model. I close this part of the paper by discussing in turn the roles of two central assumptions of the model: that there is a rigid wage for unskilled labor and that all taxes are borne by skilled workers. Each assumption is closely related to a real-life issue of considerable importance.

[13] The paradoxical possibility arises because the policy under consideration is the interdiction effort E and would disappear in the presence of an optimal tax on migrant earnings.

[14] This modification to the model is considered below.

C. *Flexible Wages*

To see the significance of a rigid wage for unskilled labor, consider now the opposite case where w is flexible. Then, in equations (2) and (3), $e=1$ and w becomes a variable. The subsequent analysis is unchanged except that w takes the role previously performed by e. In particular, the policy environment implied by the model is basically the same. The largest change has to do with the effect of a rise in E on national income, so let us take a look at that.

Differentiation of (6), now that $e=1$ and both w and v are variables, gives $(dY/dE)= L(dw/dE)+S(dv/dE)-1$. Taking note of (2) and (3), this reduces to

$$dY/dE = -I(dw/dE)-1. \tag{8}$$

That is, the welfare cost of an increased interdiction effort is the sum of the direct cost of the effort itself plus the increase in the wages paid to the initial body of illegal immigrants. This will indeed almost surely be a welfare *cost*, because there is little chance that dw/dE will be negative. Because workers are paid their marginal product, a fall in w would have to be accompanied by a *rise* in actual immigration I. But with w^* exogenous, the fall in w and rise in E would deter immigration rather than encourage it. If the host country possesses power in the international labor market, it is conceivable that w^* might fall enough to encourage more immigration. But if a fall in w^* requires a larger labor supply in source-country markets, there is little chance that this will be consistent with *both* a larger volume of successful emigration and a larger host-country apprehension ratio (due to the larger E).

The earlier skepticism about the ability of interdiction policy to raise national income is thus even stronger in the presence of a flexible wage. I leave it to the reader appropriately to rephrase the other previous results. In the rest of this paper I continue to allow a flexible wage. This makes it easier to relate my discussion to the standard theory of international trade, my point of departure. It is also (marginally) more convenient analytically. But the parallel between a flexible wage and a rigid wage is so strong that detailed separate consideration of the latter would be redundant.

D. *The Fiscal Issue*

Since border enforcement requires real resources, it must be financed. I have assumed that this is done by a tax on the earnings of skilled workers so that illegal immigrants, in particular, pay no taxes. This confronts my theory with an outstanding real-life issue: the net effect of illegal immigrants on the budget of the host-country government. It is a matter of considerable controversy, at least in the United States, whether such aliens pay taxes greater than the additional expenses that they cause. What difference does it make? To find the answer (in the world of my model at any rate), consider the alternative case in which illegal immigrants do form part of the tax base of the home country. (Since the only role of the government in my model is its enforcement activity, this amounts to specifying who pays for that enforcement). Suppose a proportional tax on production (or on value-added or on incomes) at the rate $t= E/Q$. Thus firms pass on the tax to their workers and the (post-tax) wages w and v of unskilled and skilled workers are now

$$w=(1-t)f'(u) \tag{2'}$$

$$v=(1-t)[f(u)-uf'(u)]. \tag{4'}$$

Illegal immigrants will automatically be part of the tax base. Equation (5) now becomes

$$E=tSf(u). \tag{5'}$$

As before, E directly determines w so as to equilibrate the international labor flow. Then (2′) and (5′) jointly determine t and u, which in turn imply v, from (4′), and I. But (2′) and (5′) each defines a negative relation between t and u—for any given level of E—and (2′) has no particular curvature. Thus multiple solutions are possible, and if there are multiple solutions, some of them will necessarily respond in perverse fashion to changes in E.

We are accordingly driven to the conclusion that border enforcement policy—in the form of a choice of E—is an effective means of influencing the wage paid to those native unskilled workers and legal migrants that compete directly with illegal immigrants, but that it could well be an unpredictable and unstable influence upon both the number of illegal immigrants who actually enter the country and upon the wages of native skilled workers who compete only indirectly with the migrants. By contrast, if the tax base is restricted to skilled labor, equilibrium is unique and responds in an intuitive way to changes in E.[15]

E. *Overview of Border Enforcement*

To sum up, the present model has directed attention to three issues. The *fiscal issue*, whether the migrants on balance contribute to or detract from the host-government surplus, is relevant to the ability of the authorities to use interdiction policy to control actual migration. The *income issue* concerns the effect of enforcement on the national income of the host country. This is likely negative, so that such policy is costly and so confronts the country with the problem of minimizing the cost of an enforcement policy pursued for some other reason. Finally the authorities are confronted with the *bundling problem*: how to attain simultaneously independent goals regarding the volume of immigration and the domestic distribution of income.

[15] The model is simple enough to render these conclusions reasonably transparent. An intensification of border enforcement E must raise the domestic wage received by unskilled workers. There are two ways this can come about. One is to reduce the number of unskilled workers per skilled worker, so that the marginal product of the former rises. This requires a reduction in immigration. A low elasticity of substitution between the two types of labor obviously makes it easier to produce a change in real wages in this way. The second method is to increase national product enough so that the after-tax wage received by unskilled workers rises, even if their marginal product falls. Such an increase in output can come about only if more migrants are employed. This method is facilitated by a high elasticity of substitution between the two types of workers (which limits the decline in the migrants' marginal product when more of them are employed) and by a large share of unskilled labor in national income (which limits the proportional rise in I necessary to expand output significantly). The corresponding algebra is straightforward.

II. Internal Enforcement Policy

The discussion thus far furnishes two good reasons to hunt for an additional policy tool to supplement interdiction. First, border enforcement is quite probably costly. This raises the possibility of trying to find an assortment of policies to produce the same result at lower cost. Public debate in the United States has for years been dominated by the belief that a border enforcement policy consistent with national goals would in fact prove far too costly.

The second reason is that the authorities are likely to have independent goals regarding both internal income distribution and the volume of immigration. Thus there is need for an instrument capable of unbundling these two targets.

The rest of this paper concerns the second instrument. Enforcement has thus far been assumed confined to the border so that, once inside, illegal immigrants are indistinguishable from other unskilled workers. I now consider domestic enforcement policies intended to make it relatively more difficult for illegal immigrants in the domestic marketplace. Such policies seem very promising in regard to the two reasons just discussed for looking at additional tools. First, as is suggested by the frequent public suggestions that illegal entry attempts will abate once the employment prospects of immigrants are reduced, such domestic enforcement promises to exert an effect independent from that of border enforcement. Thus a combination of such policies could potentially reduce costs. Second, if the employment of illegal immigrants is prohibited, and if this prohibition is backed up by some degree of enforcement, firms will no longer see the two types of workers as identical. Thus such a policy fosters the hope of unbundling the two goals of internal income distribution and the volume of immigration.

There are also practical reasons for considering domestic enforcement. Policy debate in the United States has for several years focused upon the advisability and consequences of the adoption of such policies. Also they are important in many northern European countries, where only modest efforts are devoted to border enforcement.

A. *Domestic Enforcement*

Suppose that domestic enforcement takes the form of random inspections of firms. Let J denote the number of illegal immigrants discovered at work by such inspections, and let D denote the total resources, in terms of domestic output, devoted to the inspection effort. I assume that

$$J/I = h(D)$$

where $h' > 0$, $h'' < 0$, $h(0) = 0$, $h < 1$. Suppose for simplicity that illegal immigrants who are caught working are paid anyway, so that only employers are subjected to direct penalties. (Thus the direct punishment from border enforcement falls exclusively upon migrants and that from domestic enforcement exclusively upon firms.) Suppose that this penalty can be expressed as an incremental cost of k^* per discovered illegal immigrant.[16] Continue to denote by w the wage received by illegal immigrants and by e the employment rate of unskilled workers. My earlier argument continues to apply, so w is still determined by border enforcement E—again, I take w^* as exogenous. Now let w_L denote the wage received by domestic unskilled workers and by legal migrants. A crucial consideration is whether domestic firms are able to distinguish these workers from illegal immigrants. If they cannot, $w = w_L$ and the effect of domestic enforcement will be to disadvantage unskilled workers relative to native skilled workers. For the firm that hires one of the former takes the risk that the new employee is an illegal immigrant and that the authorities will discover this fact, thereby increasing the expected cost to the firm of employing this worker by $k^*h(D)I/U$. But if firms can distinguish illegal immigrants, domestic enforcement will disadvantage these laborers relative to native unskilled workers and legal migrants. Thus the effect of the policy upon the welfare of this latter group is essentially sensitive to the ability of firms to distinguish between the two types of potential employees. I consider first the case where a firm can distinguish.

B. *Complete Discernment*

If firms can tell whether a prospective unskilled employee is an illegal immigrant or not, (2) will apply only to native unskilled workers and legal migrants and so is replaced by

$$w_L = f'(u). \tag{2''}$$

Risk-neutral firms will ensure that w_L equals the expected labor cost of hiring an illegal immigrant:

$$w_L = w + h(D)k^*. \tag{9}$$

Finally, taxes must now finance both enforcement policies:

$$D + E = tvS. \tag{10}$$

Thus E determines w in (1), this then gives w_L from (9), which then determines u from (2″), v from (4), and t from (10). Each policy pair (D, E) generates a unique equilibrium.[17]

[16] This paper treats as exogenous the two penalties, k and k^*. This is quite compelling as regards k, since the penalty born by aliens who fail to gain entry consists to a large degree of the opportunity cost of the time lost in being sent back. But the penalty k^* levied on employers presumably would correspond in substantial part to fines and/or prison sentences; so there is a good case for treating both D and k^* as policy variables. But to make k^* endogenous in a satisfactory way would require addressing the relation between that variable and the performance of enforcement officials, with regard to zeal, corruption, and so on. All this does not seem to me a natural and essential part of the topic of this paper, so I sidestep it by simply taking k^* as exogenous.

[17] The above argument assumes that k^* is a social cost. If it is not, in whole or in part (for example, if k^*

I wish to expose the considerations relevant to the choice of policy mix between D and E. The authorities might be using policy to try to control the level of illegal immigration I for fundamentally social reasons, or to try to influence domestic income distribution by affecting w_L. An immediate conclusion is that the mix of policy tools is irrelevant to any tradeoff between these possible targets; that is, the addition of border enforcement does not solve the unbundling problem. This is because w_L is directly related to u—and therefore to I as well—by equation (2″), regardless of the relative reliance of the authorities on border enforcement or on internal enforcement.

Consider, then, whether combining domestic enforcement with border enforcement might allow a reduction in the cost of whatever policy goal is chosen. Suppose that the authorities have as a policy target a particular level of w_L, either for its own sake or because of its associated level of I. Then substitution of this target value and of equation (1) into (9) yields

$$w_L = w^* + kg(E)/[1-g(E)] + h(D)k^*. \tag{11}$$

Equation (11) determines all combinations of the policy variables D and E that will achieve the stipulated target. These combinations are graphed as the AA' schedule in Figure 1. As drawn, the schedule has a negative slope and is strictly convex to the origin.

At point A, the wage received by illegal immigrants $w = w^*$. Moving down the schedule, with border enforcement substituted for domestic enforcement so as to keep the actual level I of illegal immigration unchanged, the wage given the latter, w, steadily rises, until at point A' it equals w_L, the (constant) wage received by native unskilled workers and legal migrants. It is also clear that such a substitution, by raising E, must increase the

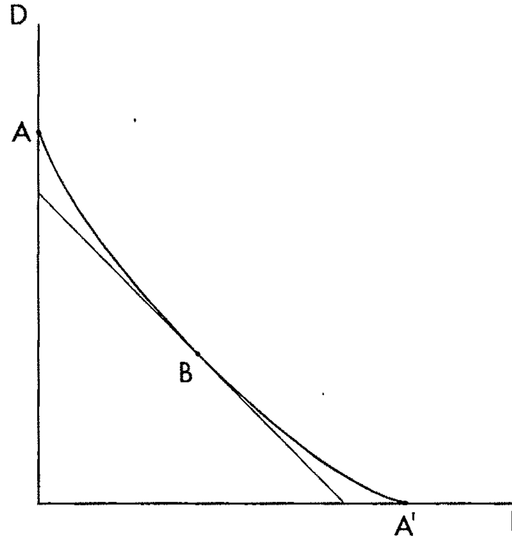


FIGURE 1

ratio C/M. With I unchanged and D falling, J must fall as well.

What determines which (D, E) combination the authorities should choose? This depends, of course, upon their objectives. To minimize the total cost $D + E$ of immigration policy, the authorities need to choose that combination of D and E for which the AA' schedule has a slope of minus unity, illustrated by point B in Figure 1.

By moving from A to B, the host country can *both* reduce its net costs *and* improve the welfare (*ex post*) of migrants illegally residing within its borders, while continuing to attain the w_L target. But beyond B, native welfare must be sacrificed in order to raise migrant welfare further. Moving toward A' will clearly raise w. The level of I will be unchanged, as will the incomes of S and L. The rise in wI must just equal, from (9), the fall in k^*J: penalties on employers are substituted by the wages of illegal immigrants.[18]

In order to find the socially optimal point on AA', it is necessary to be more specific

reflects a fine collected by the government), then the appropriate amount should be deducted from the right-hand side of (10). This does not affect the argument that each (D, E) generates a unique equilibrium.

[18]Since $D + E$ is increased by any departure from B, the after-tax incomes of skilled workers necessarily fall, with v remaining constant.

about the employer penalties k^*J. These penalties might constitute a real social cost in the form of lost output—as I have implicitly assumed thus far—or they might simply be fines which accrue to the government—in which case they should be deducted from the left-hand side of expression (10). Suppose the former. Then national income, net of immigration policy, is equal to: $Sf(u) - wI - E - D - k^*J = Sf(u) - w_L I + h(D)k^*I - E - D - k^*J = Sf(u) - w_L I - E - D$. Now $Sf(u) - w_L I$ is determined by the presumed social policy. Thus national income is maximized by minimizing $E + D$, the cost of enforcement. If, on the other hand, the employer penalties are not real costs but instead fines collected by the government, the term $-k^*J$ should not be included in the maximand. In this case maximizing national income is the same as minimizing $E + D - k^*J$. But now this latter term measures the cost of enforcement. Thus in either case this cost should be minimized in order to maximize national income.

When employer penalties constitute a real social cost, this will of course be at point B in Figure 1. If instead the penalties levied on firms take the form of fines, it will evidently be in the national interest to depart from point B in a northwest direction along AA'.

Finally, it should be pointed out that, given the policy variables D and E, changes in legal migration L would again be exactly offset by equal changes in I (and so also in J, M and C), leaving w_L and U unaltered. But, if $D > 0$, $w_L > w$ so that it is better (from the native point of view) to keep any migration illegal, that is, $L = 0$ is now optimal policy.

This analysis has, of course, proceeded on the assumption that the markets in which the illegal immigrants find themselves do in fact clear. Consider instead the unemployment model discussed earlier. Suppose that the unskilled labor wage w_L is again exogenously determined. Then the spending on internal enforcement, D, will directly determine the wage w received by employed illegal immigrants, from equation (9), but this has no effect on host-country income distribution. Again we have $w(E) = ew$, so, once D is given, spending on interdiction E directly determines the employment rate e of illegal immigrants. This is in turn linked to the volume of immigration, since $e(L + I) = Su$. Thus it remains true that the volume of immigration is bundled with domestic unskilled labor market conditions (interpreted as the employment rate, in this case). Internal enforcement offers extra control only over the national cost of the chosen policy, since variations in D produce only variations in what must be spent for immigrant labor, w.[19]

In summary, supplementing interdiction at the border with domestic enforcement does allow the authorities to achieve their chosen policy target at a lower cost than would the use of border enforcement alone. However, the basic problem that the internal distribution of income and the level of immigration are rigidly linked together remains.

C. *No Discernment*

These conclusions are based on the assumption that domestic firms are indeed able to distinguish illegal immigrants from legal unskilled workers. To get a handle on the significance of this assumption, consider briefly the opposite case. That is, assume now that firms are unable to distinguish the two.[20] Then a firm will think that there is a probability equal to $I/[L + I]$ that any unskilled worker applying for a job is in fact an illegal immigrant. As before, the expected penalty for hiring an illegal alien is $h(D)k^*$. If workers are paid the expected value, to the firm, of their marginal product, then the wage received by both legal and illegal unskilled labor must satisfy

$$(9') \quad w = f'(u) - h(D)[1 - (L/uS)]k^*.$$

[19]Note that if w is pegged, instead of the legal wage, spending on internal enforcement D will directly influence the latter, so that the bundling problem will be solved. But this description of the labor market seems less intuitively appealing then the one in the text. In any case, discriminating between the two would require a detailed analysis quite beyond the scope of the present paper.

[20]Since this case is of interest only if the government can distinguish illegal aliens when it inspects firms, it is necessary that the government have access to some means of verification not available to private firms.

Since the immigrant wage must still be given by $w(E)$, this implies that (2) must now be replaced by

$$(2''') \quad w(E) = f'(u) - h(D)[1-(L/uS)]k^*.$$

It is immediate from (2‴) that the addition of a second policy tool has now unbundled the internal distribution of income and the level of illegal immigration. Furthermore, the instruments should be assigned to the two targets in a very specific way. Border enforcement E will determine directly the unskilled labor wage, w, irrespective of domestic enforcement D. The latter can accordingly then be adjusted to cause u (and therefore I) to assume its desired value.[21]

Notice the peculiar position in which legal unskilled workers find themselves. On the one hand, the fact that firms cannot distinguish them from illegal aliens means they must share the latter's fate: both are disadvantaged relative to skilled workers. But it is just this which breaks the link between w and I, and so allows the authorities free rein to use interdiction policy to influence w, without concern for the affect of such policy on the volume of immigration.[22]

D. *Endogenous Discernment*

I've looked at the two extreme cases where firms either completely distinguish legal unskilled workers from illegal aliens, or completely fail to make such a distinction. Consider now the more realistic situation where limited discernment takes place. This will not merely constitute a mix of the two cases just considered, but will introduce distinctive features of its own, especially when account is taken of the fact that the degree of discernment will be at least partially endogenous.

If firms cannot distinguish between illegal immigrants and other unskilled workers, domestic enforcement will disadvantage unskilled workers (including native unskilled workers and legal immigrants) relative to skilled native workers. Thus legitimate unskilled workers have an incentive to adopt measures to distinguish themselves from their illegal rivals, and to support public policies designed to produce such distinctions.

It is not surprising, then, that the policies should be central to public debate on immigration enforcement in host countries. In an analysis of the issues they raise, two aspects seem crucial: the effort illegal immigrants devote to trying to pass themselves off as legal, and the effects on those legal residents who might be mistaken for illegal entrants.

Suppose that a proposed policy attempts to distinguish illegal migrants from legitimate unskilled workers. This presumably involves "doing something" to the latter, since the former would not cooperate with such an effort. The policy could be one of issuing identification cards to legal workers.

The unskilled labor force consists, as before, of illegal immigrants I plus legal unskilled workers L who would never be mistaken for illegal entrants, but now there is also a group N of legal unskilled workers who could be so mistaken. Thus the total unskilled labor force is $U = L + N + I$. Let p denote the probability that a member of I succeeds in passing himself off as a member of N. Presumably attaining any positive value of p requires the migrant to spend some amount μ on deceptive efforts (such as the purchase of counterfeit identification cards):

$$\mu = H(p)$$

where $H(0) = 0$, $0 < p < 1$, $H' > 0$, $H'' > 0$. For simplicity it is assumed that H is of constant elasticity $e_D = pH'/\mu$, the "elasticity of deception."

[21] Note that I am here interpreting the internal distribution target as concern about the wage of unskilled workers. The wage of skilled workers will still depend upon u, and the taxes they pay will depend upon $D+E$.

[22] The unemployment variation of this model is again qualitatively similar. Interdiction policy E directly determines the employment rate e of unskilled labor from the relation $w(E) = ew$, where w is the pegged wage. Equation (9′) shows how domestic enforcement policy D directly determines u. Since $L + I = Su/e$, D can be used to control the volume of immigration independently of the employment rate implemented by the choice of E.

As before, a potential migrant who attempts entry faces the probability g of failing, and thus earning $(w^* - k)$, and the probability $(1-g)$ of gaining entry. In the latter event, he spends μ on deception and faces the conditional probability p of successfully passing himself off as a member of N and thereby earning their wage w_N, and the conditional probability $(1-p)$ of being recognized as a member of I and thereby earning their wage, w. Assuming as before that attempted migration M adjusts to equate the expected reward of attempting entry to that of staying behind,

$$(12)\quad w^* = (w^* - k)g + [w_N p + w(1-p) - \mu](1-g).$$

Illegal entrants will choose μ so as to maximize their expected earnings, $w_N p + w(1-p) - \mu$. Setting the derivative of this expression equal to zero gives the condition

$$(13)\quad w_N - w = \mu\varepsilon_D / p.$$

The second-order condition is $\varepsilon_D > 1$, which I henceforth assume. A boundary solution, with p equal to either 0 or 1, is also possible. If $p = 0$, illegal immigrants in fact make no attempts at deception, and members of N are treated just like members of L: my earlier analysis of discriminating firms remains relevant. If, on the other hand, $p = 1$, members of N find themselves treated just like members of I, with $w_N = w(E) + \mu$. Since these boundary cases are easily understood, I now assume an interior solution.

Workers recognized as illegal immigrants will, as before, be paid the value of their marginal product adjusted for the chance that they will be discovered and their employers penalized:

$$(14)\quad w + h(D)k^* = f'(u).$$

Substituting (13) and (14) into (12) then gives

$$f'(u) - h(D)k^* = w(E) + (1-\varepsilon_D)\mu$$

where $w(E)$ is as defined previously in (1). Rearranging this term and recalling the definition of u gives

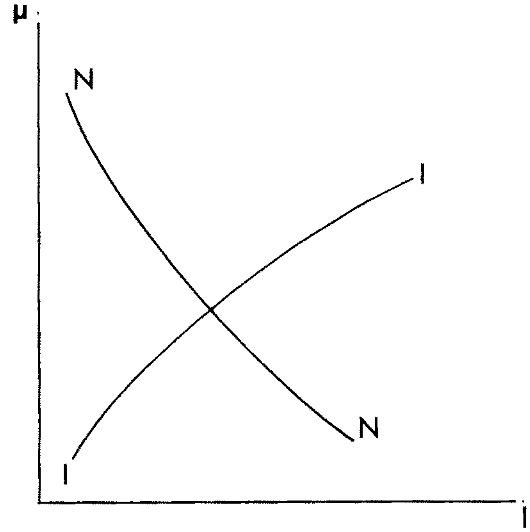


FIGURE 2

$$(15)\quad f'([N + I + L]/S) - (1-\varepsilon_D)\mu = w(E) + h(D)k^*.$$

This expression is graphed as the II schedule in Figure 2. It depicts those combinations of illegal entrants I and deceptive efforts μ which cause the illegal labor market to clear. Everywhere along this schedule the numbers attempting to migrate are just such as to equilibrate the expected earnings from migration to the earnings from staying, successful entrants make the optimal deceptive effort, and those entrants recognized by employers for what they are receive the value of their marginal product less the expected employer penalty.

Clearly the II schedule will have a positive slope, as depicted in Figure 2, since the elasticity of deception exceeds unity. Note also that equation (15) can be written as $w = w(E) + (1-\varepsilon_D)\mu$. Recall that $w(E)$ is the wage that illegal immigrants would receive were there no possibility of deception, and w is the wage paid those perceived to be illegal. Thus migrants known to be illegal entrants are paid less than they would receive with the same border enforcement

effort, were deception impossible. For those migrants this is a secondary burden, since they also pay μ for the deception attempt. For their employers it is a benefit, since they pay less for the illegal labor they hire.

Next consider the markets for legal unskilled workers. Those who are clearly legal (L) will receive a wage w_L equal to the value of their marginal product, as described in (2′). But a member of N could, from the viewpoint of a prospective employer, actually be an illegal immigrant whose employment would bring with it the possibility of a penalty. The chance that an apparent member of N is actually a member of I is $pI/(N+pI)$, and the expected penalty from employing an illegal alien is hk^*. Thus members of N earn the wage satisfying

$$(16)\quad w_N+[pI/(N+pI)]h(D)k^*=f'(u).$$

Subtracting (16) from (14) and using (13) gives

$$(17)\quad [(pI+N)/N]\mu/p=h(D)k^*/\varepsilon_D.$$

This relation, graphed as the NN schedule in Figure 2, shows all combinations of I and μ that clear the market for legal unskilled workers apparently of type N. Clearly the NN schedule must have a negative slope, as depicted in Figure 2, since $\varepsilon_D>1$.

It is evident from (15) and (17) that an intensification of border enforcement E will shift the II schedule to the left while leaving the NN schedule untouched. Thus an increase in E will reduce the volume of illegal immigration I, but cause those entrants to devote more effort to deception. This means that the perceived volume of illegal immigration, $(1-p)I$, will fall by more than will I itself: the intensification of border enforcement will appear to be more successful than it really is (and than it would appear to be if there were no policy of internal enforcement).

An increase in domestic enforcement D will also shift the II schedule to the left, while in addition shifting NN to the right. Thus the rise in D has an ambiguous effect on illegal immigration but does significantly increase attempts at deception. Thus such a policy shift is likely to produce a large decline in perceived immigration, and thus appear to be quite successful, even if in fact it has little effect on I!

To proceed further, let us examine analytically the effects of policy changes. Differentiating the expressions for the II and NN schedules (equations (15) and (17)) with respect to E and solving yields

$$\hat{\mu}=-ew[w(E)/N\Delta]\hat{E}$$

and

$$\hat{I}=\varepsilon_w[w(E)/N\Delta]\times[(N(\varepsilon_D-1)+pI\varepsilon_D)/pI\varepsilon_D]\hat{E}.$$

In these expressions, ε_w denotes the elasticity of $w(E)$, $[w'E/w(E)$, and

$$\Delta=-[\theta_S w_L/eU][(N(\varepsilon_D-1)+pI\varepsilon_D)/pN\varepsilon_D]-(\varepsilon_D-1)\mu/N<0,$$

where θ_S denotes the distributive share of skilled labor. That is, Δ is the Jacobian determinant of the system (15),(17):

$$\Delta=\begin{bmatrix}[-\theta_S w_L]/eU & \varepsilon_D-1\\ \mu/N & [N(\varepsilon_D-1)\\ & \quad +pI\varepsilon_D]/pN\varepsilon_D\end{bmatrix}.$$

These expressions confirm that an increase in border enforcement E will lower illegal immigration I and intensify attempted deception μ. The net effect on the volume of aliens successfully passing themselves off as legitimate unskilled workers, pI, is then given by

$$\hat{p}+\hat{I}=(\hat{\mu}/\varepsilon_D)+\hat{I}$$
$$=\varepsilon_w[w(E)/N\Delta][(\varepsilon_D-1)/\varepsilon_D]\times[(pI+N)/pI]\hat{E}.$$

Thus pI must fall as a result of the rise in border enforcement: the intensified deception effort does not make up for the reduction in illegal immigration.

We can now see how the increased border enforcement affects the various labor groups. The fall in I implies a fall in U and therefore in u as well; this implies, from (2′) and (14), that the wages of native unskilled workers of unambiguous identity, w_L, and those of illegal migrants recognized as such, w, both increase, and in identical amounts. Because of the fall in pI, the wages of members of N (and of those illegal migrants passing themselves off as legitimate) increases even more than w_L and w. (This conclusion is by no means obvious; it depends critically upon $\varepsilon_D > 1$—the second-order condition for optimal deception.) The losers are native skilled workers and those migrants who fail to gain entry as a result of the rise in E. The basic conclusion, then, is that it is the members of that group susceptible to being confused with illegal migrants that stands to gain the most from an intensification of border enforcement, in spite of the fact that one of the consequences of such an intensification will be to induce illegal entrants to redouble their efforts to pass themselves off as members of the susceptible group.

An intensification of internal enforcement in the percentage $\hat{D}$ produces the following results.

$$\hat{\mu} = -[(pI+N)/pN][\varepsilon_I/\Delta] \times [(\theta_S w_L/eU)+(\varepsilon_D\mu/N)]\hat{D}$$

$$\hat{I} = (hk^*\varepsilon_I/\Delta)([(1-p)/pI] \times [(\varepsilon_D-1)/e_D]+(1/N))\hat{D}$$

where $\varepsilon_I = h'D/h$, the elasticity of the internal enforcement schedule.

It follows from these expressions that an intensification of internal enforcement increases the deception effort μ, as indicated by the geometry, but also unambiguously reduces the volume of illegal immigration I. Thus the wages of clearly legal unskilled workers (L) rise as a result of the increase in D. However the effect on pI is ambiguous, and therefore so also is the effect on the wage (w_N) of legal workers who might be mistaken for illegal immigrants. Indeed it could be the case that the welfare of members of N might actually be improved by an increase in internal enforcement efforts! Clearly illegal immigrants are also affected ambiguously; in (14) both the value of these worker's marginal product, $f'(u)$, and the unexpected penalty of employing them, $h(D)k^*$, rise. Obviously these workers are disadvantaged relative to the clearly legal, with $w_L - w$ rising in the proportion $eI\hat{D}$, from (14). They are also disadvantaged relative to the ambiguous group: it follows from (13) that $w_N - w$ rises in the proportion $[(\varepsilon_D-1)/\varepsilon_D]\hat{\mu}$.

An intensification of either border enforcement or internal enforcement will raise the wages of those workers liable to be mistaken for illegal immigrants relative to those who clearly are illegal immigrants. But whereas more border enforcement definitely makes the former group better off both absolutely and relative to clearly legal unskilled workers, increased internal enforcement need not do so.

D. *Unemployment and Endogenous Discernment*

As before, there is a parallel analysis featuring rigid wages and unemployment. I will describe the model and its equilibrium, but leave to the reader the task of repeating the above exercises in the new context.

Suppose, again, that the wage w paid to workers recognized to be illegal immigrants is exogenously determined, that jobs are allocated among these workers on the basis of a random draw, and that e denotes the employed fraction of all workers known to be illegal (not of all workers actually illegal). The wages w_L and w_N continue to be flexible. Then ew replaces w in equations (12) and (13), which can be substituted into each other to yield

$$ew = w(E) - \mu(\varepsilon_D - 1). \tag{18}$$

This is the analog to the II schedule. For each level E of border enforcement, it gives a positive relation between the employment rate e in the illegal immigrant labor market and the degree μ of deceptive effort. Note that this relation is invariant with respect to D.

With w rigid, equation (14) gives u as a function of D. Equation (16) is unchanged. Actual employment of workers believed to be members of N equals $N+pI$, and the actual employment of workers known to be present illegally equals $e(1-p)I$, so

$$u=[L+N+pI+e(1-p)I]/S.$$

Thus we have $N/[N+pI]=[p+e(1-p)]/[p([uS-L]/N)+e(1-p)]=\Omega(p,e,u)$.

Substituting (14), (16), and (13) into each other—and recalling that ew replaces w in the latter—gives

$$\text{(19)}\quad \mu\varepsilon_D/p+(e-1)w = h(D)k^*\Omega(p,e,u).$$

Since u is determined directly by D, equation (19) gives another relation between e and μ, for each value of D. This is the analog to the earlier NN schedule. Note that its position is independent of E, but its slope is not determinate from our assumptions.

Equilibrium is once again determined by the intersection of the II and NN schedules, that is, by the simultaneous solution of (18) and (19). The same comparative statics exercises as before can be performed in this analogous framework.

III. Conclusions

The more important implications of this paper appear to be the following.

Border enforcement policy is an effective means of controlling the unskilled labor employment rate, or, in the flexible wage version of the model, the wage of unskilled labor.

However, if illegal immigrants form part of the tax base, such interdiction could well be an unpredictable and unstable influence upon both the number of illegal immigrants who actually enter the country and upon the wages of native skilled workers who compete only indirectly with the migrants.

With a given interdiction policy, varying the number of immigrants admitted legally will have no effect on the number who actually enter.

Border enforcement policy will probably reduce national income, even if the country has power in the international labor market.

The wage of unskilled workers (or their employment rate) and the volume of illegal immigration are likely to be distinct policy targets, but they are linked together technologically and cannot be unbundled with border enforcement policy.

Domestic enforcement policies will disadvantage illegal aliens relative to legal workers if firms can distinguish the former, but if they cannot, such policies will harm all unskilled workers relative to skilled labor.

A country can reduce the cost of its immigration policy by employing a mixture of border and domestic measures rather than relying on just one type of enforcement.

Domestic enforcement cannot unbundle the unskilled labor wage and the volume of immigration if firms distinguish illegal aliens, but it can do so if they do not distinguish.

If illegal aliens can take measures to improve their chances of passing themselves off as legal residents, both border enforcement and, especially, domestic enforcement will appear to be more successful than they actually are.

An intensification of border enforcement will be especially attractive to those legal workers liable to be confused with illegal aliens. The use of domestic enforcement allows the authorities to unbundle the wage of this group from the volume of immigration.

REFERENCES

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, March/April 1968, *76*, 169–217.

Bhagwati, Jagdish N., "Alternative Theories of Illegal Trade: Economic Consequences and Statistical Detection," *Weltwirtschaftliches Archiv*, 1981, *117*, 409–27.

______, "International Migration of the Highly Skilled: Economics, Ethics and Taxes," *Third World Quarterly*, July 1979, *1*, 17–30.

______ **and Hamada, Koichi**, "The Brain Drain,

International Integration of Markets for Professionals and Unemployment: A Theoretical Analysis," *Journal of Development Economics*, June 1974, *1*, 19–42.

______ **and Hansen, Bent,** "A Theoretical Analysis of Smuggling," *Quarterly Journal of Economics*, May 1973, *87*, 172–87.

Djajic, Slobodan, "Illegal Aliens, Unemployment and Immigration Policy," unpublished manuscript, 1985.

Ehrlich, Isaac, "Participation in Illegitimate Activities: An Economic Analysis," in G. S. Becker and W. M. Landes, eds., *Essays in the Economics of Crime and Punishment*, New York: Columbia University Press, 1974.

Ethier, Wilfred. J., "International Trade and Labor Migration," *American Economic Review*, September 1985, *75*, 691–707.

______, "International Trade Theory and International Migration," Discussion Paper No. 5, Migration and Development Program, Harvard University, January 1984.

Glazer, Nathan, *Clamor at the Gates: The New American Immigration*, San Francisco: Institute for Contemporary Studies, 1985

Harris, John R. and Todaro, Michael, "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review*, March 1970, *60*, 126–42.

Pitt, Mark M., "Smuggling and Price Disparity," *Journal of International Economics*, November 1981, *11*, 447–58.

Rodriguez, Carlos A., "Brain Drain and Economic Growth: A Dynamic Model," *Journal of Development Economics*, September 1975, *2*, 223–247.

Rational and Self-Fulfilling Balance-of-Payments Crises

By MAURICE OBSTFELD*

The collapse of a fixed exchange rate is typically marked by one or more balance-of-payments crises in which speculators acquire a large portion of the central bank's foreign reserves as the bank attempts in vain to support its currency. Economists have long attributed such crises to inappropriate domestic policies that ultimately place the central bank in the uncomfortable position of offering speculators a one-sided bet. The recent balance-of-payments literature, beginning with the work of Paul Krugman (1979), tends to support the foregoing view. When asset holders have perfect foresight and the current peg must eventually be abandoned, profit maximization dictates that a sharp attack on the central bank's reserves occur at some point on the economy's path.[1] Speculative attacks appear to be self-fulfilling, since they may occur even when the level of reserves seems sufficient to handle "normal" balance-of-payments deficits. But under the view sketched above, the attacks are inevitable, and represent an entirely rational market response to persistently conflicting internal and external macroeconomic targets.

This paper demonstrates the existence of circumstances in which balance-of-payments crises may indeed be purely self-fulfilling events rather than the inevitable result of unsustainable macroeconomic policies. Such crises are apparently unnecessary and collapse an exchange rate that would otherwise have been viable. They reflect not irrational private behavior, but an indeterminacy of equilibrium that may arise when agents expect a speculative attack to cause a sharp change in government macroeconomic policies.

Section I sets out a simple stochastic fixed exchange rate model in which (a) domestic credit may deviate from its constant mean level only by a serially correlated disturbance with finite variance, (b) shocks to domestic credit are small in a well-defined sense, and (c) agents do not expect any change in domestic-credit policy in the event of a balance-of-payments crisis. In that setting, a run on the central bank's foreign reserves is a probability-zero event. The section goes on to discuss the timing of the exchange rate's inevitable collapse when domestic credit follows a random walk with drift rather than a stationary, finite-variance process.[2] Section II then changes assumption (c) by postulating that agents expect an exchange rate collapse to set off an inflationary domestic-credit policy. Under the new assumption, self-

*Department of Economics, Columbia University, New York, NY 10027. José Saúl Lizondo made detailed and extremely useful comments on an earlier draft of this paper. I am grateful also for helpful discussions with Guillermo Calvo, Robert Flood, and Peter Garber. Any errors and all opinions are mine. The National Science Foundation and the Alfred P. Sloan Foundation provided financial support.

[1] The literature on rational balance-of-payments crises derives from Stephen Salant and Dale Henderson (1978), who discuss the concept of a rational speculative attack in the context of the gold market. Salant (1983) provides further developments. The analogy between external payments crises and attacks in resource markets is closest in the continuous-time, nonstochastic setting explored by Krugman, by Robert Flood and Peter Garber (1984a, Section II), and my article (1984), among others. The essential reason is that it is only in continuous time that an anticipated discrete exchange rate jump generally entails "abnormal" profit opportunities. My paper (1986) presents a discrete-time-maximizing model in which the collapse of the exchange rate may involve two successive speculative attacks: one on the first day of floating, and one the period before. In stochastic models such as those described below the notion of "speculative attack" becomes even more blurred if one defines speculation as any reduction in private money demand in anticipation of the fixed rate's possible collapse. In this sense, there may be any number of "attacks" before the rate finally does collapse.

[2] Flood and Garber (1984a, Section III) have already studied this problem, but their discussion is incomplete at one point. Because the logical step they omit is critical for understanding the self-fulfilling equilibria of Section II, I discuss it at length in Section I.

fulfilling runs become a possibility. Indeed, the economy is shown to possess a continuum of equilibria, each corresponding to a different subjective assessment of the likelihood of an exchange rate collapse. The nominal interest rate in this economy will at times exceed the world rate; and a positive innovation in domestic credit will widen any international interest differential, even while the exchange rate remains fixed. Section III places the self-fulfilling crisis example in the context of related literature on bank runs, bubbles, and extrinsic uncertainty. The Appendix discusses some technical questions that arise in the text.

I. The Economics of Rational Crises

The model employed is the simple linear one developed by Robert Flood and Peter Garber (1984a) and used in my 1984 article. Under the assumptions made in those papers (and in the earlier literature), speculative attacks on the currency occur in a setting where the eventual abandonment of the current fixed exchange rate is inevitable. Speculation can never occur if market conditions are consistent with the unconditional and indefinite maintenance of the fixed exchange rate. After briefly setting out the model, this section explains the economics of those results. The next section will give an example of a scenario under which even a perpetually viable exchange rate can be attacked.

A small country enjoys perfect international capital mobility, and its residents consume and produce a single good. If S_t is the domestic-currency price of foreign exchange and P_t^* is the foreign-currency price of output, the domestic price level P_t is given by $S_t P_t^*$, for all times t. If i_t^* is the nominal interest rate on foreign-currency securities, the domestic nominal interest rate i_t is given by $i_t = i_t^* + E_t[(S_{t+1}/S_t) - 1]$, where $E_t[\cdot]$ denotes an expectation conditional on time t information. Both P^* and i^* are assumed to be constants, equal to 1 and 0, respectively. Let $\bar{S}$ denote the level of the initially fixed exchange rate.

Domestic money is held by domestic residents only, and consists entirely of the liabilities of the central bank. Let R_t be the book value of central bank foreign reserves and D_t domestic credit. Equilibrium in this economy is determined by the equality of money demand,

$$(1) \qquad M_t^d / P_t = \alpha - \beta i_t,$$

and money supply,

$$(2) \qquad M_t^s = R_t + D_t.$$

Assume first that as long as the exchange rate is fixed, domestic credit evolves according to the law

$$(3) \qquad D_t = \bar{D} + v_t,$$

where the disturbance v_t follows the covariance-stationary $AR(1)$ process

$$(4) \quad v_t = \rho v_{t-1} + \varepsilon_t (0 \leq \rho < 1, \ E_{t-1}[\varepsilon_t] = 0)$$

and the innovations ε_t are serially independent. The assumptions made above imply that if the exchange rate is fixed at $\bar{S}$ and expected to remain so next period, equilibrium reserves are $R_t = \alpha \bar{S} - D_t$. In this case, the domestic interest rate i_t and the world rate i_t^* coincide.

The exchange rate regime collapses on the date that private domestic wealth owners acquire the entire remaining stock of central bank foreign reserves.[3] A collapse clearly presupposes the existence of some lower bound on central bank reserves, $\bar{R}$. It is worth noting, though, that there is no reason in principle why a central bank facing a perfect international capital market cannot borrow *indefinitely* to support the exchange rate, provided it raises taxes to service the external debt it incurs (see my 1986 paper). Nonetheless, $\bar{R}$ is assumed to be an exogenous, possibly negative, constant for the purpose of the present analysis. It is assumed further that mean reserves under a perma-

[3] The central bank may commit only a portion of its reserves to the defense of the exchange rate, as in Krugman and my article (1984). For the present analysis, I assume this is not the case.

nently fixed rate exceed $\bar{R}$, that is, that $\alpha\bar{S} - \bar{D} - \bar{R} > 0$.

In the models of Krugman, Flood and Garber (1984a), and myself (1984, 1986), a steadily growing domestic-credit stock makes a breakdown of the fixed-rate regime inevitable. (I return to this result shortly.) But the alternative credit-growth process given by (3) and (4) does not imply that the fixed exchange rate is unconditionally viable. A large enough realization of the random variable ε_t (given v_{t-1}) can clearly drive reserves to their limit $\bar{R}$, forcing an unexpected abandonment of the fixed rate $\bar{S}$ and a depreciation of the currency. It is convenient to rule out this possibility by assuming that $\text{Prob}[\varepsilon < (1-\rho)(\alpha\bar{S} - \bar{D} - \bar{R})] = 1$. (This implies, by (4), that $\text{Prob}[v < \alpha\bar{S} - \bar{D} - \bar{R}] = 1$.) Under this additional assumption the fixed exchange rate can, with probability one, persist indefinitely.

If the domestic-credit rule described by (3) and (4) is followed regardless of the exchange rate regime, domestic credit does not grow steadily and an exchange rate collapse is a probability-zero event. To see why this is so, recall first that equilibrium reserves always exceed $\bar{R}$ if the exchange rate is pegged at $\bar{S}$ and expected to remain pegged. Suppose now that there is an equilibrium for the economy involving the private acquisition of the entire official reserve stock on some date T. I will show that, with probability one, the currency *appreciates* (i.e., S falls) as the central bank is forced to withdraw from the foreign exchange market. This implies that the hypothesized collapse cannot occur along an equilibrium path of the economy. Any investor who anticipates the exchange rate's path will not participate in the attack (even if he believes everyone else will), but will prefer to wait until others have dislodged the exchange rate from its peg so that he can buy foreign assets at a lower domestic-currency price. Because no one will wish to participate in the attack, it cannot occur at time T.

To see that the floating rate $\tilde{S}_T$ resulting from a run at time T lies below $\bar{S}$, solve for the economy's rational expectations equilibrium under free floating. All official reserves in excess of $\bar{R}$ have been acquired by speculators when floating commences. Together with the assumed international parity conditions, (1) and (2) therefore imply that the floating rate's evolution is governed by the difference equation

$$(5)\quad -\beta E_t[\tilde{S}_{t+1}] + (\alpha+\beta)\tilde{S}_t = \bar{R} + D_t \qquad (t \geq T).$$

The saddle-path solution for $\tilde{S}_T$ is[4]

$$(6)\quad \tilde{S}_T = (\alpha+\beta)^{-1} \times \sum_{j=0}^{\infty} \left(\frac{\beta}{\alpha+\beta}\right)^j E_T[\bar{R} + D_{T+j}].$$

Under assumptions (3) and (4), $\tilde{S}_T$ can be written

$$\tilde{S}_T = \alpha^{-1}(\bar{R} + \bar{D}) + [\alpha + \beta(1-\rho)]^{-1} v_T.$$

Because $\bar{S} > \alpha^{-1}(\bar{R} + \bar{D})$, $\tilde{S}_T < \bar{S}$ if $v_T \leq 0$. It was assumed above that $v_T < \alpha\bar{S} - \bar{D} - \bar{R}$ with probability one. It therefore follows from the solution for $\tilde{S}_T$ that if $v_T > 0$, $\tilde{S}_T < \alpha^{-1}(\bar{R} + \bar{D} + v_T) < \bar{S}$ with probability one. Since the currency must appreciate immediately if the central bank loses all reserves at T, equilibrium attacks are probability-zero events. A consequence is that the domestic and foreign interest rates must always coincide in the present setting.

The key element in the above argument was the stipulation that no abrupt policy change is expected to occur as the result of a crisis: the domestic-credit process described by (3) and (4) does not change, and the central bank simply withdraws from the foreign exchange market when its reserves are

[4]The general solution to the expectational difference equation (5) is the saddle-path solution (6) plus a term of the form $K\xi_t((\alpha+\beta)/\beta)^t$, where K is an arbitrary constant and $\{\xi_t\}$ is any stochastic process such that $E_t[\xi_{t+1}] = \xi_t$. The saddle-path solution excludes self-fulfilling divergent bubbles by imposing the initial condition $K = 0$. My papers with Kenneth Rogoff (1983, 1986) provide justification for the saddle-path condition in the context of an optimizing monetary model. See Section III, below, for further discussion.

exhausted. Once the foregoing stipulation is dropped, the equilibrium of the economy may become indeterminate. As a result, attacks become possible, even if the exchange rate would have been viable forever in their absence.

Before demonstrating this possibility in the next section, it will be useful to describe how an exchange rate collapse occurs in the present model if domestic credit grows monotonically over time. This is the scenario studied by Flood and Garber (1984a). Consider an economy in which domestic credit evolves according to the rule

$$D_t = D_{t-1} + \mu_t (E_{t-1}[\mu_t] = \mu > 0), \tag{7}$$

where $\text{Prob}[\mu_t \geq 0] = 1$, for all t. In terms of (3) and (4), ρ assumes the value 1 and the mean of ε shifts upward. It is assumed again that (7) holds regardless of the exchange rate regime, and that no foreign exchange intervention occurs after a collapse.

If the exchange rate floats at any time t (and if central bank reserves accordingly are constant at their lower limit $\bar{R}$), the equilibrium rate $\tilde{S}_t$ is again given by (6). Under domestic-credit rule (7), however,

$$\tilde{S}_t = \alpha^{-1}(\bar{R} + D_t) + \alpha^{-2}\beta\mu. \tag{8}$$

Suppose first that $\tilde{S}_t \leq \bar{S}$. As in the previous discussion, the exchange rate cannot be floating in period t. If the equilibrium floating rate when reserves equal $\bar{R}$ is below $\bar{S}$, reserves would exceed $\bar{R}$ in an equilibrium with the exchange rate pegged at $\bar{S}$. This would be true even if the exchange rate were expected to float in $t+1$, since the expected depreciation between t and $t+1$ would be reduced by a rise in the rate from $\tilde{S}_t$ to $\bar{S}$. Thus the central bank can certainly peg the rate at $\bar{S}$ through period t unless the public acquires the reserve stock in an attack. But no individual would find it profitable to join in an attack that causes an instantaneous exchange rate appreciation. (The bank will lose its remaining reserves in an attack if $\tilde{S}_t = \bar{S}$, but the exchange rate remains at $\bar{S}$ until the next period. I therefore consider it "nonfloating" on date t, a matter of definition.)

What if $\tilde{S}_t > \bar{S}$? Flood and Garber (1984a) note correctly that the exchange rate must float in the first period that this inequality holds. However, the rationale they offer for this assertion—that the condition $\tilde{S}_t > \bar{S}$ offers speculators an opportunity to profit at official expense—merely shows (recall the previous paragraph) that *if* a run is expected, it will pay for all to join in. In other words, while a path for the economy such that the exchange rate collapses the first time $\tilde{S}_t > \bar{S}$ is an equilibrium path, it remains to show that no other outcome is consistent with intertemporal equilibrium.

This is easily done once a precise definition of "equilibrium" is adopted. At any time t the state of the economy is defined by the current realization of domestic credit D_t. Following Stephen Salant (1983), define an equilibrium exchange rate function $S(D_t)$ by the properties:

(a) for all realizations D_t, there is a reserve level $R_t \geq \bar{R}$ such that

$$R_t + D_t = \alpha S(D_t) - \beta E_t[S(D_{t+1}) - S(D_t)];$$

(b) $S(D_t) = \bar{S}$ if $R_t > \bar{R}$.

Then under the central bank behavior assumed above, the following result can be demonstrated:

THEOREM 1: *If $S(D_t)$ is an equilibrium exchange rate function, there is a critical domestic-credit level $\tilde{D}$ such that $S(D_t) > \bar{S}$ if and only if $D_t > \tilde{D}$. The threshold $\tilde{D}$ is defined by*

$$\tilde{D} \equiv \inf\{D: \alpha^{-1}(\bar{R} + D) + \alpha^{-2}\beta\mu > \bar{S}\}.$$

It follows from (8) that $S(D_t) = \tilde{S}_t$ whenever $\tilde{S}_t > \bar{S}$.

PROOF:

It has already been shown that $S(D_t) = \bar{S}$ in equilibrium if $D_t \leq \tilde{D}$ (so that $\tilde{S}_t \leq \bar{S}$). To prove that $S(D_t) > \bar{S}$ whenever $D_t > \tilde{D}$ (in which case $R_t = \bar{R}$ and $S(D_t) = \tilde{S}_t$), assume the contrary. Thus, let $\tilde{D}' > \tilde{D}$ be the smallest level of domestic credit such that $S(\tilde{D}') > \bar{S}$ (more precisely, the infimum of the set of D with $S(D) > \bar{S}$; this set is nonempty because reserves are limited). I

will show that there are realizations of domestic credit D_t between $\tilde{D}$ and $\tilde{D}'$ such that $S(D_t)=\bar{S}$ is inconsistent with money market equilibrium.

So let $D_t \in (\tilde{D}, \tilde{D}')$. Since $R_t \geq \bar{R}$, money market equilibrium requires

$$(9)\quad \bar{R}+D_t \leq \alpha\bar{S}-\beta E_t[S(D_{t+1})-\bar{S}]$$

if $S(D_t)=\bar{S}$. Define $\delta(D_t) \equiv \mathrm{Prob}[\mu_{t+1} \geq \tilde{D}' - D_t]$. Then

$$\begin{aligned} E_t[S(D_{t+1})] &= [1-\delta(D_t)]\bar{S} \\ &\quad + \delta(D_t)E_t[\tilde{S}_{t+1}|\mu_{t+1} \geq \tilde{D}'-D_t] \\ &= [1-\delta(D_t)]\bar{S}+\delta(D_t)\{\alpha^{-1}(\bar{R}+D_t) \\ &\quad + \alpha^{-1}E[\mu_{t+1}|\mu_{t+1} \geq \tilde{D}'-D_t]+\alpha^{-2}\beta\mu\}. \end{aligned}$$

Let $g(\cdot)$ denote the probability density function for μ_{t+1}. Then the necessary condition (9) can be written

$$(10)\quad \tilde{S}_t \leq \bar{S}+(\beta/\alpha[\alpha+\beta\delta(D_t)]) \times \int_0^{\tilde{D}'-D_t} \mu_{t+1}g(\mu_{t+1})\,d\mu_{t+1}.$$

As Figure 1 shows, the left-hand side of (10) rises linearly with D_t, and $\tilde{S}_t \geq \bar{S}$ for $D_t = \tilde{D}$. The right-hand side of (10) is nonincreasing, reaching its minimal value $\bar{S}$ when $D_t = \tilde{D}'$. Finally, note that no exchange rate function with $S(D_t)=\bar{S}$ for $D_t < \tilde{D}'$ defines an equilibrium because condition (10) is violated when $D_t \in (D', \tilde{D}')$.[5]

It is clear that the form of domestic-credit rule (7) was a key ingredient in the above demonstration that the exchange rate must float as soon as $\tilde{S}_t$ exceeds $\bar{S}$. The observation that everyone would wish to participate in a run if one were expected does not itself prove that a run must occur. The next section underlines this point by providing an example of an economy in which the exchange rate can remain fixed even if its floating rate "shadow" value $\tilde{S}_t$ exceeds $\bar{S}$.

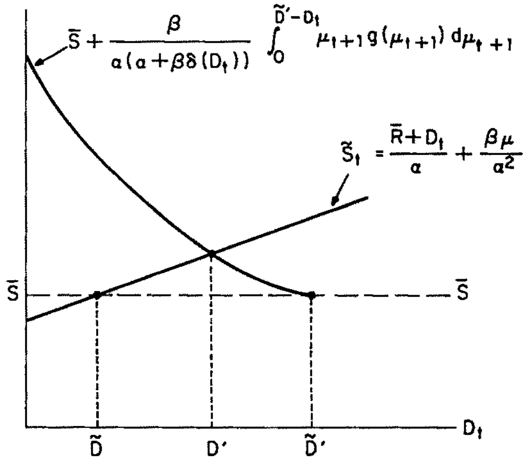


FIGURE 1

II. Self-Fulfilling Crises: An Example

This section gives an example of a crisis-induced policy change which, if anticipated, leads to indeterminacy of equilibrium and the possibility of self-fulfilling speculative attacks on a fixed exchange rate. The example involves the expected adoption of an inflationary domestic-credit growth rule in the event the fixed exchange rate collapses. Initially, however, the exchange rate is fixed at $\bar{S}$ and the domestic-credit process is again described by (3) and (4), together with the restriction on the range of the domestic-credit innovation ε.

Assume now that the public holds the following expectation: if a collapse occurs at any time T, the central bank allows the exchange rate to float forever and switches to the domestic-credit growth rule (7).

There are a number of stories that might buttress this assumption of an expected regime change. For example, when reserves hit $\bar{R}$, the central bank can no longer borrow

[5] Figure 1 incorporates the assumption that the density function $g(\cdot)$ is continuous; but the proof is easily extended to encompass discrete distributions. Using the techniques in Salant, a theorem like the one in the text could be proved for credit processes more general than (7). However, this would require the assumption that the central bank pegs the exchange rate at $\bar{S}$ whenever possible. The restriction that $\mathrm{Prob}[\mu_t \geq 0]=1$, also imposed by Flood and Garber (1984a), makes the latter assumption unnecessary.

externally, and the government may need to resort to inflationary finance.[6] This need is exacerbated by the sharp rise in official net foreign indebtedness caused by the loss of reserves to private speculators. The recent experiences of Argentina and Chile confirm that financial disorder and capital flight may foreshadow heightened future inflation (see, for example, Carlos Diaz-Alejandro, 1986). While the one-way causation postulated here is inadequate as a description of those countries' experiences, the example serves to elucidate a factor that may have played a role.[7]

One equilibrium path for the economy is the one described in the preceding section. If the public expects that no collapse will ever occur, there is no run on the central bank's foreign reserves and no switch in the domestic-credit process. Expectations are self-fulfilling, and the fixed-rate regime remains in place forever with probability one.

I now argue that there are infinitely many alternative equilibria, each corresponding to a different set of public beliefs about the probability of a run. These, too, are self-fulfilling equilibria. Under the policy scenario assumed here, the authorities are expected to validate any run *ex post* by shifting to an inflationary policy. The argument of the previous section, which ruled out crises in the policy environment described by unconditional adherence to (3) and (4), no longer applies.

As a first step in the construction of an alternative equilibrium, note that in the event of an exchange rate collapse at time T, the value of the floating exchange rate is again given by (8), which is reproduced here with the substitution $D_T = \bar{D} + v_T$:

$$\tilde{S}_T = \alpha^{-1}(\bar{R} + \bar{D} + v_T) + \alpha^{-2}\beta\mu.$$

Because a crisis now induces expectations of future inflation, it is possible that the exchange rate depreciates if the central bank is forced to leave the foreign exchange market. This occurs if $\tilde{S}_T > \bar{S}$, that is, if

$$\alpha^{-1}(\bar{R} + \bar{D} + v_T) + \alpha^{-2}\beta\mu > \bar{S}.$$

Clearly the foregoing inequality can hold even when the equilibrium reserve level in the no-run equilibrium, $\alpha\bar{S} - \bar{D} - v_T$, exceeds $\bar{R}$; it implies that a run may take place whenever

$$v_T > (\alpha\bar{S} - \bar{R} - \bar{D}) - \alpha^{-1}\beta\mu \equiv \bar{C}. \tag{11}$$

Because $\mu > 0$, inequality (11) may hold even under the boundedness assumption on ε made in Section I.

Consider now the following sequence of events. At the start of a period T, the value of current domestic credit is revealed. Simultaneously, an exogenous lottery determines the state of nature. There are two possible states. Private agents believe that in state 1, which occurs with probability π, a run on the central bank's reserves will take place if, and only if, $v_T > \bar{C}$ (inequality (11)). But they believe that in state 2, which occurs with probability $1 - \pi$, no run will take place. If state 1 occurs and $v_T > \bar{C}$, every agent will find it advantageous to participate in the expected run by selling the central bank as much domestic money as possible. Because the exchange rate is expected to depreciate once the central bank can no longer peg it at $\bar{S}$, agents will flee the domestic currency to avoid a sure capital loss.

The distribution of events described above clearly defines a stochastic equilibrium in the special case $\pi = 0$, namely, the no-run equilibrium of the previous section. If positive choices of π also define equilibria, runs can occur, and the domestic interest rate i can lie above the world rate i^* even while the exchange rate is fixed. This would be necessary to compensate domestic bond holders for

[6] The evaporation of the government's credit lines does not by itself imply that the private sector can no longer lend abroad or repatriate foreign assets. Thus, the home interest rate may remain linked to i^* by interest parity. This would not be the case if the government were to impose capital controls in response to an exchange rate crisis. Charles Wyplosz (1983) examines the role of controls in crises.

[7] Other examples are possible. Multiple equilibria also arise if agents expect a run to provoke an immediate discrete devaluation, but expect no devaluation otherwise. See my paper (1984) for a related discussion of devaluation.

capital losses expected in the event of a collapse.[8]

Do there exist equilibria with $\pi > 0$? The technical obstacle to resolving this question is connected with the foregoing observation concerning the domestic interest rate i. If $\pi > 0$, shocks to domestic credit may alter the domestic interest rate, and thus need not lead to equal offsetting changes in reserves. If a positive domestic credit innovation raises i (a conjecture that will be verified below), the upper bound on ε assumed in Section I no longer prevents unexpectedly high domestic credit creation from driving reserves to $\bar{R}$ and wiping out the fixed-rate regime without warning. If inequality (11) can hold *only* for shocks that drive reserves below $\bar{R}$, the notion of fixed exchange rate equilibria with $\pi > 0$ is vacuous.

To avoid this problem, it is sufficient to restrict the range of ε more tightly. Assume there exists a positive $\bar{B}$ satisfying $\bar{C} < \bar{B} < \alpha\bar{S} - \bar{D} - \bar{R}$ such that $\mathrm{Prob}[\varepsilon < (1-\rho)\bar{B}] = 1$. As is shown in the Appendix, there exists a $\bar{\pi} > 0$ so small that for $\pi < \bar{\pi}$, reserves are no less than $\bar{R}$ in equilibrium if state 1 has never occurred and the exchange rate is fixed at $\bar{S}$. The crisis probabilities π in the interval $[0, \bar{\pi})$ define a continuum of possible stochastic equilibria for the economy in which the exchange rate is fixed if state 1 has never occurred when $v_t > \bar{C}$.

With this more stringent restriction on ε, it is straightforward to compute the domestic interest rate for π in $[0, \bar{\pi})$ while the exchange rate is fixed. Let $D_t = \bar{D} + v_t$ summarize the state of the economy on a date t when no run occurs. If $v_{t+1} < \bar{C}$, a crisis cannot occur next period, so S_{t+1} will remain at $\bar{S}$. If $v_{t+1} > \bar{C}$, a crisis will occur with probability π.[9] Define $q(v_t) \equiv \mathrm{Prob}[\varepsilon_{t+1} > \bar{C} - \rho v_t]$. Then by (8),

$$E_t[S_{t+1}] = [1 - \pi q(v_t)]\bar{S} + \pi q(v_t)\{\alpha^{-1}(\bar{R} + \bar{D} + \rho v_t + E[\varepsilon_{t+1} | \varepsilon_{t+1} > \bar{C} - \rho v_t]) + \alpha^{-2}\beta\mu\}. \tag{12}$$

Equation (12) states that the exchange rate expected to prevail next period (given that no run has yet occurred) is a weighted sum of $\bar{S}$ and the rate expected to equilibrate asset markets in the event of an exchange rate collapse. As the latter cannot be lower than the peg $\bar{S}$, $E_t[S_{t+1}] \geq \bar{S}$ and so $i_t = i_t^* + E_t[(S_{t+1}/\bar{S}) - 1] \geq i_t^*$.[10]

Because domestic credit shocks are serially correlated, a higher value of v_t (given that there is no run) raises the probability that v_{t+1} will exceed the critical level $\bar{C}$ at which a run becomes possible. In fact, an increase in v_t causes a rise in next period's expected exchange rate, and thus a rise in the interest rate i_t. To see this, let $f(\cdot)$ denote the (continuous) probability density function of the random variable ε. Then

$$q(v_t) = \int_{\bar{C} - \rho v_t}^{(1-\rho)\bar{B}} f(\varepsilon)\, d\varepsilon$$

$$(\text{or } 0, \text{ if } \bar{C} - \rho v_t \geq (1-\rho)\bar{B})$$

and

$$E[\varepsilon_{t+1} | \varepsilon_{t+1} \geq \bar{C} - \rho v_t] = \int_{\bar{C} - \rho v_t}^{(1-\rho)\bar{B}} (\varepsilon f(\varepsilon)/q(v_t))\, d\varepsilon \ (\text{if } q(v_t) > 0)$$

so that differentiation of (12) implies

$$d(E_t[S_{t+1}])/dv_t = \alpha^{-1}\pi\rho q(v_t) \geq 0, \tag{13}$$

for any π. A consequence of (13) is that a positive domestic-credit shock may actually reduce money demand, occasioning a more-than-offsetting reserve loss.

[8] Guillermo Calvo (1983) discusses the possibility that an expected devaluation of uncertain timing can lead to rises in the domestic nominal interest rate and the *ex post* real interest rate. José Saúl Lizondo (1983) relates the behavior of foreign exchange futures prices under fixed exchange rates to similar considerations. Of course, in the Flood and Garber (1984a) model discussed above, i exceeds i^* whenever there is a chance that the fixed exchange rate will collapse next period.

[9] It is being assumed that the borderline case $\varepsilon_{t+1} = \bar{C} - \rho v_t$ has a zero probability weight.

[10] Given the assumed upper bound on the possible realizations of ε, it is possible that v_t is so small that $q(v_t) = 0$. In this case $E_t[S_{t+1}] = \bar{S}$ and $i_t = i_t^*$. Inequality (13) below remains valid, however.

III. Discussion

The previous section presented an example of an economy in which self-fulfilling expectations give rise to a continuum of possible equilibria. Even though a crisis is not inevitable, agents believe that the central bank will respond to crises by embarking on a program of heightened inflation. The belief that the authorities will (in effect) ratify crises makes it unprofitable for any individual speculator to hold domestic currency while a run is taking place.

In this context, balance-of-payments crises are very similar to bank runs. Douglas Diamond and Philip Dybvig (1983) present a stylized model of financial intermediation in which there are two equilibria: one in which agents have confidence in the solvency of financial intermediaries, and one in which lack of confidence leads to a run. Both equilibria involve self-fulfilling expectations because banks fail if, and only if, there is a run. As Section I showed, balance-of-payments crises, unlike bank runs, need not be self-ratifying. The stability of a pegged-rate regime hinges on the anticipated response of the authorities.

If runs are to be made possible in the model of Section II, it is necessary to endow agents with rational subjective probabilities of runs. This was accomplished by randomizing over the run and no-run equilibria. Olivier Blanchard (1979) uses this device to construct an example of a nonstationary asset-market bubble that "crashes" with probability one. He assumes that when the asset price is on a bubble path, there is a time-invariant probability that it will return to its fundamental or saddle-path value next period. The type of uncertainty determining the equilibrium agents believe will prevail has been labelled "extrinsic" uncertainty by David Cass and Karl Shell (1983). (See also Costas Azariadis, 1981.) Those authors study the allocative effects of extrinsic uncertainty in a utility-maximizing model with restricted market participation.

In their study of gold monetization, Flood and Garber (1984b) give an example of a self-fulfilling run similar to the one explored above. Their paper shows how an otherwise viable gold standard might break down if agents anticipate that it will collapse some time in the future. As in the example of self-fulfilling balance-of-payments crises, and for the same reason, indeterminacy arises only when agents expect the authorities to resort to inflationary finance in the wake of a regime collapse. While Flood and Garber's gold model is nonstochastic, the introduction of extrinsic uncertainty would result in additional equilibrium paths.

Flood and Garber (1984a) suggest that the timing of a run may be indeterminate for reasons different from those explored above. If the floating exchange rate prevailing after a run can reflect divergent speculative bubbles as well as market fundamentals, the equilibrium floating rate is indeterminate. This implies that runs can in principle occur at any time. The example studied in this paper explicitly assumes that the floating exchange rate depends entirely on its fundamental determinants, however (equation (6)).[11] Rogoff and I (1983, 1986), using a maximizing model show how the government can prevent such bubbles by using its fiscal powers to guarantee a minimal real redemption value for money. These results are applicable also to the stochastic divergent bubbles of the type studied by Blanchard and others.

The fractional backing described by myself and Rogoff (1983, 1986) is an example of an official guarantee which, though never exercised, precludes inefficient equilibria supported by self-fulfilling expectations. The deposit insurance scheme studied by Diamond and Dybvig is another example in this class.[12] These papers make the important argument that anticipated government policies can block the emergence of certain suboptimal competitive equilibria. In a sense, the present

[11] In proving Theorem 1, it was assumed that the equilibrium exchange rate is a function of domestic credit alone. This amounts to excluding from consideration nonstationary bubbles and any form of extrinsic uncertainty.

[12] Note, however, that while the Diamond-Dybvig insurance scheme requires essentially full backing of deposits, the policy that precludes explosive price-level bubbles in my papers with Rogoff works for arbitrarily small amounts of real currency backing.

paper turns that argument on its head. Expected government actions may lead to undesired outcomes in economies that would function more efficiently otherwise.

Appendix

This Appendix demonstrates that the crisis probability π can always be chosen so small that if the first "state of nature" described in Section II has not occurred, noncollapse of the fixed-rate regime is consistent with money market equilibrium.

In order that this be true, equilibrium reserves in state 2 (the no-run state of nature) must always exceed $\overline{R}$ for sufficiently small π. By equations (1)–(3), this means that when $E_t[S_{t+1}]$ is given by (12),

$$\text{(A1)}\quad R_t = \alpha\overline{S} - \overline{D} - v_t - \beta(E_t[S_{t+1}] - \overline{S}) > \overline{R}$$

for all t. I argue that there exists a positive $\overline{\pi}$ such that (A1) always holds for $\pi < \overline{\pi}$.

Let $E_t[\tilde{S}_{t+1}|\varepsilon_{t+1} > \overline{C} - \rho v_t]$ denote the equilibrium floating exchange rate expected to materialize on date $t+1$ if a run occurs. By (12), (A1) can be written

$$\text{(A2)}\quad \alpha\overline{S} - \overline{D} - v_t - \beta\pi q(v_t) \times\left(E_t[\tilde{S}_{t+1}|\varepsilon_{t+1} > \overline{C} - \rho v_t] - \overline{S}\right) > \overline{R}.$$

Equation (12) and inequality (13) imply that $q(v_t)(E_t[\tilde{S}_{t+1}|\varepsilon_{t+1} > \overline{C} - \rho v_t] - \overline{S})$ can be written as a function of v_t, $\Phi(\cdot)$, where $\Phi'(\cdot) \geq 0$. With this definition, (A2) becomes

$$\text{(A3)}\quad \alpha\overline{S} - \overline{D} - \overline{R} > v_t + \beta\pi\Phi(v_t).$$

Recall that ε is bounded above by $(1-\rho)\overline{B}$; it follows that v_t can never exceed $\overline{B}$. Since $\Phi(\cdot)$ is nondecreasing, it attains its maximum possible value at $v_t = \overline{B}$. Let $\overline{\pi}$ be the positive solution to

$$\alpha\overline{S} - \overline{D} - \overline{R} - \overline{B} - \beta\overline{\pi}\Phi(\overline{B}) = 0.$$

It is now clear that for any $\pi < \overline{\pi}$, (A3) holds with probability 1. Thus if $\pi < \overline{\pi}$, the fixed-rate regime is always consistent with money market equilibrium.

References

Azariadis, Costas, "Self-Fulfilling Prophecies," *Journal of Economic Theory*, December 1981, *25*, 380–96.

Blanchard, Olivier Jean, "Speculative Bubbles, Crashes and Rational Expectations," *Economics Letters*, 1979, *3*, 387–89.

Calvo, Guillermo A., "Trying to Stabilize: Some Theoretical Reflections Based on the Case of Argentina," in Pedro Aspe Armella et al., eds., *Financial Policies and the World Capital Market: The Problem of Latin American Countries*. Chicago: University of Chicago Press, 1983.

Cass, David, and Shell, Karl, "Do Sunspots Matter?," *Journal of Political Economy*, April 1983, *91*, 193–227.

Diamond, Douglas W., and Dybvig, Philip H., "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, June 1983, *91*, 401–19.

Diaz-Alejandro, Carlos F., "Goodbye Financial Repression, Hello Financial Crash," *Journal of Development Economics*, 1986, forthcoming.

Flood, Robert P. and Garber, Peter M., (1984a) "Collapsing Exchange Rate Regimes: Some Linear Examples," *Journal of International Economics*, August 1984, *17*, 1–14.

______ **and** ______, (1984b) "Gold Monetization and Gold Discipline," *Journal of Political Economy*, February 1984, *92*, 90–107.

Krugman, Paul, "A Model of Balance-of-Payments Crises," *Journal of Money, Credit, and Banking*, August 1979, *11*, 311–25.

Lizondo, José Saúl, "Foreign Exchange Futures Prices under Fixed Exchange Rates," *Journal of International Economics*, February 1983, *14*, 69–84.

Obstfeld, Maurice, "Balance-of-Payments Crises and Devaluation," *Journal of Money, Credit, and Banking*, May 1984, *16*, 208–17.

______, "Speculative Attack and the External Constraint in a Maximizing Model of

the Balance of Payments," *Canadian Journal of Economics*, 1986, forthcoming.

______ **and Rogoff, Kenneth,** "Speculative Hyperinflations in Maximizing Models: Can We Rule Them Out?," *Journal of Political Economy*, August 1983, *91*, 675–87.

______ **and** ______, "Ruling Out Divergent Speculative Bubbles," *Journal of Monetary Economics*, 1986, forthcoming.

Salant, Stephen W., "The Vulnerability of Price Stabilization Schemes to Speculative Attack," *Journal of Political Economy* February 1983, *91*, 1–38.

______ **and Henderson, Dale W.,** "Market Anticipations of Government Policies and the Price of Gold," *Journal of Political Economy*, August 1978, *86*, 627–48.

Wyplosz, Charles, "Capital Controls and Balance of Payments Crises," mimeo., National Bureau of Economic Research, 1983.

Employment, Hours, and Earnings in the Depression: An Analysis of Eight Manufacturing Industries

By BEN S. BERNANKE*

Seismologists learn more from one large earthquake than from a dozen small tremors. On the same principle, the Great Depression of the 1930's would appear to present an important opportunity for the study of the effects of business cycles on the labor market. In no other period for which we have data do output, labor input, and labor compensation exhibit such severe short-run variations.

Despite this apparent opportunity, modern econometric analyses of labor markets have typically made little use of pre-World War II data. There are some significant exceptions. In the class of papers that assume continuous labor market equilibrium, the best known example is by Robert Lucas and Leonard Rapping (1969). This influential piece was followed by Michael Darby (1976), who basically supported the Lucas-Rapping approach, and by Joseph Altonji and Orley Ashenfelter (1980) and Altonji (1982), who were critical of it. Among papers that allow for market disequilibrium, work by Harvey Rosen and Richard Quandt (1978) and Ashenfelter (1980) should be noted.[1] However, none of the papers cited, I think it is fair to say, is the definitive study of 1930's labor markets. They have in common at least two deficiencies in this regard.

First, all of this work has employed annual and highly aggregated data. This reflects the fact that none of the papers is focused on the 1930's per se but include prewar data only as part of a longer-period study. Since none of the papers uses data from before 1929, any conclusion drawn about the prewar period is based on at most a dozen or so observations.

Second, the papers are limited in their capacity to rationalize the movements of a number of key labor market time-series. For example, none of them addresses the radical fluctuations in the length of the workweek which occurred during the depression, a phenomenon which the present research will argue is a quite important part of the overall story. As is documented in my forthcoming paper with James Powell and my working paper (1985), variations in the workweek contributed nearly as much as did changes in employment to the overall variance in labor input during this period (in the manufacturing industries studied)—in contrast to the postwar period, during which employment change was the much more important factor. Workweek reductions were also surprisingly persistent: in the iron and steel industry, hours of work (which were about fifty-five hours per week during the late 1920's) did not average as much as forty hours weekly in any year from 1932 to 1939, and for long periods were considerably less.

Perhaps more significant, and more puzzling, than the behavior of the workweek, was the behavior of the real wage. My paper with Powell showed, for the industry data set used also in this paper, that real wages were typically countercyclical during the prewar period. This countercyclicality is equally apparent if indexes of wage rates[2] are used instead of average hourly earnings to measure real wages; it seems to have held for the manufacturing sector as a whole (Alan

*Department of Economics and Woodrow Wilson School, Princeton, Princeton NJ 08542. I thank participants in workshops at Stanford, MIT, Harvard, Chicago, Carnegie-Mellon, Rochester, Princeton, and Pennsylvania for comments on the first draft of this paper. Numerous colleagues were also helpful. The Center for Economic Policy Research and the Hoover Institution provided support.

[1] Martin N. Baily (1983) gives an interesting discussion of labor markets in the 1930's, but does not estimate a structural econometric model.

[2] The data on wage rates, available for the first six industries and through August 1931 only, are from Daniel Creamer (1950).

Stockman, 1983) as well as for individual industries. The tendency of real wages to rise despite high unemployment was especially striking during the major depression cycle (1929–37): real wages rose during the initial downturn (1930–31). They rose sharply again in 1933–34 and 1937, despite unemployment rates of 20.9 percent in 1933, 16.2 percent in 1934, and 9.2 percent in 1937 (according to Darby's correction of Stanley Lebergott's 1964 figures). In contrast, my paper with Powell found some evidence of real wage procyclicality in similar data for the postwar period.

Why real wages should rise when the demand for labor is presumably very low[3] is difficult for any existing approach, equilibrium or disequilibrium, to explain: On the equilibrium side, Lucas and Rapping (1972) admitted in a reply to Albert Rees (1970) that their model could not explain the relation of wages and employment for the period from 1933 until the war; they did claim success for 1929–33. However, as Rees (1972) noted in his rejoinder, even this more restricted claim requires that the negative effects of falling nominal wages and prices on labor supply in 1929–33 strongly dominate the positive effect of the steadily rising real wage.

How could deflation have reduced labor supply even though real wages were rising? The original Lucas-Rapping explanation appears to be that falling nominal wages and prices depressed current labor supply by raising workers' expectations of inflation (expectations are assumed to be adaptive in the log of the price level) and lowering *ex ante* real interest rates. In light of Lucas (1972), an alternative rationale for this effect of deflation is that workers mistakenly interpreted the fall in money wages as a (local?) decline in real wages. The first explanation is hard to maintain, both on quantitative grounds and also given that, *ex post*, real interest rates in 1930–33 were the highest of the century. The second explanation relies on an extremely slow diffusion of information about wages and prices. In either case, it seems unlikely that the impact of falling nominal wages and prices would be strong enough and persistent enough to explain the data.[4]

The disequilibrium, or Keynesian, explanation for the behavior of real wages in the 1930's (at least in 1930–33) is that nominal wage "stickiness" and the sharp deflation combined to create an unplanned increase in real wages; higher real wages forced firms up their labor demand curves, adding to unemployment.[5] Now it cannot be denied that money wages are more inertial than prices (in the sense that they exhibit less high-frequency variation), although the economic interpretation of this fact is in dispute. Indeed, the present paper will conclude that the inertia of nominal wages must be given some role in the explanation of real wage behavior. However, the problem with the Keynesian story as a complete explanation, in my view, is the *degree* of unexplained wage rigidity that must be accepted in order to fit this model to the facts. For example, for their 1930–73 sample, Rosen and Quandt estimated that up to four years are required to eliminate *half* of an initial discrepancy between the actual wage and its equilibrium path;[6] presumably, the same model estimated on prewar data would yield smaller if not negative speeds of real wage adjustment.

[3] Throughout I will maintain the premise (with which I believe most economists would agree) that prewar business cycles were characterized primarily by fluctuations in the aggregate volume of labor demanded, rather than in the volume of labor supplied. It is, of course, not difficult to explain countercyclicality in the real wage when labor supply is fluctuating.

[4] Darby estimates an equilibrium model that does better than the model of Lucas and Rapping in explaining the 1930's. This model is discussed and reinterpreted in Section III below.

[5] The Keynesian story does not have a very satisfactory answer to why firms prefer laying off workers to cutting the wage, although some theoretical attempts (relying, for example, on adverse-selection problems) have been made in that direction. The sticky-wage story is also not very useful for explaining 1933–39, when real wages rose despite high unemployment and *rising* prices.

[6] Rosen and Quandt's 1978 paper postulated sticky real (rather than nominal) wages. In their 1985 work, they estimated a sophisticated disequilibrium model in which sticky nominal wages are assumed; they again found very slow speeds of wage adjustment.

Such slow rates of adjustment are incompatible with what we know of the institutions and practices prevalent in most prewar labor markets.[7]

The present paper gives a new empirical analysis of depression-era labor markets, with particular attention to rectifying the two problems just cited. First, instead of aggregate annual data, I employ monthly data for each of eight manufacturing industries. I also extend the sample period back to 1923, which gives more than 200 time-series observations for each industry. This previously unexploited data set (described in more detail in Section II and in the Appendix) appears to be a rich source of information.

Second, to search for an improved explanation of the behavior of labor market variables over this period, in this paper I depart from the standard equilibrium and Keynesian models in favor of a different and somewhat eclectic approach. The basis of my analysis is a model in which, as in Lucas (1970), firms may vary the number of hours each employee works per period (the intensive margin) as well as the number of people employed (the extensive margin). In combination with other, more conventional elements (including slow, but not glacial, adjustment of nominal wages to price changes), this model is able to deliver a reasonably successful explanation of the behavior of workweeks, real wages, and other important variables such as employment. The basic model may also be of independent theoretical and empirical interest; see Section I, Part G.

A caveat to the above is that this paper focuses on the labor market, not on the economy on the whole; thus, the offered "explanations" of labor market variables are only partial, in that they take the paths of industry outputs as given. The partial equilibrium approach was adopted for theoretical and econometric simplicity; it also has the advantage of producing results which are not dependent on a specific explanation of prewar fluctuations in aggregate demand. (However, I note here my view that it was the monetary and financial collapse of 1930–33 that gave the depression its unusually severe character; see my 1983 paper.)

The paper is organized as follows: Section I introduces a simple model of the labor market which builds on elements of Lucas (1970). An empirical analysis which uses this model as the starting point, but also incorporates a number of additional features, is discussed in Section II. Section III considers an alternative, more dynamic specification of the supply side of the model.

I. The Supply and Demand for Workers and Hours of Work: A Model

The model which contributes the basic elements of the analysis of this paper is described in a short article by Lucas (1970). The distinctive feature of Lucas's setup is that he assumes that firms can vary not only the number of workers (and of machines) that they employ, but also the number of hours per period in which the workers (and machines) work, that is, the "workweek." In equilibrium in this model, the manner in which changes in total labor input are divided between variations in workweeks and changes in employment depends on the nature of the production function and on worker preferences. Since, as has been noted, large fluctuations in workers' weekly hours were a prominent feature of depression labor markets, the explicit determination of workweeks in Lucas's analysis makes his model (or a related one) a natural candidate for use in the present context. As a bonus, Lucas showed that his model places no restriction

[7]In particular, for most sectors during the prewar period (including manufacturing, the subject of this study), barriers to rapid wage adjustment following economic shocks were much lower than they are today. Factors that gave firms a relatively free hand with respect to wages (and with respect to the employment relationship in general) included: the quiescence of the labor movement between the early 1920's and the latter New Deal; the fall in average skill levels which followed the introduction of mass production techniques; the ample supplies of unskilled and low-skilled workers; the low level of government intervention in labor relations; and the lack of a social consensus about the nature of workers' rights. See my 1985 paper for further discussion and references. Also recommended is the excellent book by Irving Bernstein (1960).

on the cyclical behavior of (average) real wages; thus this model is also (at least) not inconsistent with the observed countercyclical pattern of wages. Indeed, it will be shown in the analysis below that conditions that promote cyclical sensitivity of the workweek may also increase the tendency toward real wage countercyclicality. Thus, there appears to be a previously unsuspected connection between the puzzling aspects of the two time-series.

In what follows I set out a simple, static model in the spirit of Lucas's paper.[8] This model, in conjunction with some additional elements (including elementary dynamics), is the basis for the estimation reported below.

A. *Setting*

Since my data are for individual manufacturing industries, for concreteness in what follows I will consider the supply and demand of labor for a "primary" (manufacturing) sector. Each primary sector is to be thought of as being surrounded by its own "secondary" or alternative sector, in which people work at agriculture, trade, or services, or are not formally employed. The demand for the output of primary sectors is assumed to be more cyclically sensitive than the demand for secondary-sector output. (This assumption appears to be reasonable for most manufacturing industries.) Primary sectors are also assumed to be separated on some dimension (geographical or otherwise) from other primary sectors and thus do not compete directly with each other for workers. (This last assumption seems to be realistic for the 1930's; while there was much movement of workers between manufacturing and the secondary sector, few workers moved from one manufacturing sector to another. See, for example, E. Wight Bakke, 1940, p. 242.)

To reemphasize: discussions below of the supply or demand for labor refer *only* to the primary sector. The secondary, less cyclical sector is not explicitly modeled.

B. *The Supply Side*

In this model I shall be concerned with the determination of both 1) the length of the workweek, and 2) the number of workers employed, not just the total number of hours worked. Thus, on the supply side, I shall have to consider both the willingness of the individual worker to increase hours of work *and* the sensitivity of the participation rate to the returns available in the primary sector. Let us first examine the supply the hours of work by an individual worker, worker i. I will characterize the individual's supply curve of hours indirectly, through a function describing his reservation level of earnings for each level of hours worked. (This is the analogous construct to Lucas's 1970 wage schedule, $w(s)$.) Let E_{it} be the nominal earnings received by worker i in period t, H_{it} be the number of hours worked by i in t, and θ_{it} be a set of unspecified exogenous indicators known to worker i in t. Let COL_t be the period t cost of living, which will be assumed for now to be public knowledge. Now define the *earnings function*

$$(1) \qquad E_{it}(H_{it}, COL_t, \theta_{it})$$

to be the minimum (nominal) earnings necessary to induce worker i to work H_{it} hours (in the primary sector) in period t, given the cost of living COL_t and indicators θ_{it}.

I have begun by introducing the earnings function to emphasize that it is a very general concept, well-defined for almost any specification of the worker's preferences and environment. However, I will here make a number of restrictive assumptions in order to derive the earnings function for a specific, particularly simple case. I assume first that the worker has a temporally separable utility function, with within-period utility

$$(2) \qquad U_i = U_i(C_{it}, \overline{H} - H_{it}),$$

where C is consumption and $\overline{H} - H$, the complement of hours of work, is leisure. I assume also that the worker cannot borrow

[8]It should be made clear that Lucas is not to be implicated for the details of what follows, which differ substantially from his paper. Yakir Plessner and Shlomo Yitzhaki (1983) employ a model similar to that below.

or lend, but simply consumes his earnings each period ($C_{it}=E_{it}/COL_t$). With these two assumptions I rule out some complexities that occur when workers can intertemporally substitute consumption and leisure (but see Section III below). Finally, suppose that the worker has a reservation level of utility U_{it}^*, which he is able to obtain in the secondary or alternative sector. (Here, U_{it}^* is the datum affecting the worker's labor supply, i.e., $\theta_{it}=U_{it}^*$.) In this case the earnings function can be constructed period by period; it is defined by

$$(3)\qquad U_i\left(E_{it}(H_{it},COL_t,U_{it}^*)/COL_t,\bar{H}-H_{it}\right)=U_{it}^*$$

for $H_{it}>0$; $E_{it}=0$, otherwise. That is, the earnings function is an indifference curve in (E,H) space. With normal curvature assumptions on the utility function, (3) implies that the earnings function will be increasing and convex in hours, as well as increasing in the reservation level of utility. (See Figure 1.)

An important feature of the earnings function defined in (3) is that there is a discontinuity at zero hours: no payment is required to induce zero hours, but the earnings function is positive as hours approach zero from the right. The implicit assumption underlying this feature is that a worker who works any positive amount of hours in the primary sector must leave the secondary sector completely, that is, there is no moonlighting. Although the existence of the jump at zero hours is important for obtaining countercyclical real wages, it should be emphasized that the no-moonlighting assumption is much stronger than I need. With moonlighting, the earnings function will take lower values for small H than is suggested by (3), because workers will be able to make use of the extra time; however, as long as there is any fixed cost associated with moving between jobs, or simply a cost of going to work, the discontinuity at zero in the earnings function will exist.

Consider now the second component of labor supply, the supply of individual workers (i.e., the primary sector participation rate). I model labor supply to the primary sector as

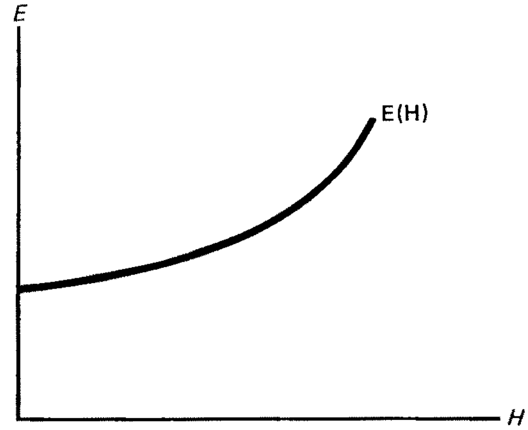


FIGURE 1. THE EARNINGS FUNCTION

increasing in the utility level offered by that sector. This can be motivated simply as follows. Assume that workers are alike in their productivity and in their utility functions, but that they differ in their secondary sector opportunities (or in other factors that affect reservation utility, such as dislike for primary-sector types of work). Specifically, let

$$(4)\qquad U_{it}^*=\gamma_t\Omega_i,$$

where γ_t is a purely time-dependent scalar and Ω_i is an individual-specific constant. The distribution function of Ω_i in the population is $G(\Omega_i)$; $G(0)=0$ and $G(\infty)=\bar{N}$, where $\bar{N}$ is the total population of potential workers. Assume that the reservation utilities of individual workers are private information, so that workers must be treated identically.[9] Then, if the primary sector wishes to employ N workers, it must provide each worker with a utility equal to at least $\gamma_t G^{-1}(N)$. Alternatively, the supply curve of workers in period t, $N_t^s(U_t^*,\gamma_t)$, can now be defined by

$$(5)\qquad N_t^s(U_t^*,\gamma_t)=G(U_t^*/\gamma_t).$$

[9]Since workers have identical utility functions and productivity, there is no opportunity for firms to induce self-selection among workers, as in Andrew Weiss (1980).

The total cost to the primary sector of employing N workers for H hours each in period t can be written

$$(6) \qquad NE(H, COL_t, \gamma_t G^{-1}(N)).$$

Per worker earnings E can now be seen to depend positively on the level of primary-sector employment N and the index of alternative opportunities γ, as well as on hours of work H and the cost of living COL.[10]

C. *The Demand Side*

I now examine the behavior of the representative firm in the primary sector, firm j. The price of the firm's output is taken as given;[11] thus, to calculate the firm's derived demand for labor, I need only to specify the production function.

The usual specification of the production function assumes that employment and the number of hours each employee works enter multiplicatively, for example, as

$$(7) \qquad Q_{jt} = F(L_{jt}, X_{jt}),$$

where Q_{jt} is the output of firm j in t, L_{jt} is total worker-hours (i.e., $L_{jt} = N_{jt}H_{jt}$, where N_{jt} is firm employment and H_{jt} is the length of the workweek), and X_{jt} is a vector of nonlabor inputs. However, as Martin Feldstein (1967) and Sherwin Rosen (1968) have noted, the assumption that employment and hours worked enter multiplicatively may not be a good one. For example, lengthening the workweek by a given percentage may affect output differently than increasing the number of workers by the same percentage.[12] Since here I particularly want to focus on the distinction between hours of work and the number of workers, I follow Feldstein in specifying the production function as

$$(8) \qquad Q_{jt} = F(N_{jt}, H_{jt}, X_{jt}).$$

This is more general than (7) if the assumption is maintained that (say, for technological reasons) each worker in the firm has a workweek of the same length.

The profit-maximization problem for firm j can be written

$$(9) \qquad \max_{\{N, H, X\}} pF(N_j, H_j, X_j) - N_j E(H_j, COL, U^*) - r(X_j),$$

where p is the output price, $r(X_j)$ is the cost of X_j, and the time subscripts are suppressed. The reservation utility of the marginal worker, U^*, depends on sectoral employment N, not firm employment N_j, and is parametric to the firm; its determination will be discussed in a moment.

The relevant first-order conditions are

$$(10) \qquad pF_N = E$$

$$(11) \qquad pF_H = N_j E_H$$

where now the capitalized subscripts denote differentiation (with respect to firm-specific variables) and the notation has been abbreviated further, in the obvious way. Equation (10) says that the firm should hire extra workers up to the point that their marginal

[10] The expression for total labor cost (6) assumes that primary-sector firms pay workers just enough to make the marginal worker indifferent between the primary and secondary sector. An alternative assumption, suggested by the "efficiency wage hypothesis" (see, for example, Janet Yellen, 1984), is that firms avoid the costs of continuous monitoring of employees by paying more than the minimum required earnings (thus giving employed workers a surplus), then firing workers caught shirking in random "spot checks." This alternative assumption, which could easily be incorporated into the present framework, has the advantage of being able to explain such phenomena as the long queues at employment offices and the extreme reluctance of the employed to leave their jobs.

[11] That is, firms are assumed to be competitive in output markets. This is admittedly not such a good assumption for some of the industries studied. See the discussion of the model simulations below.

[12] Lengthening the workweek may have diminishing returns because of increased worker fatigue; increased employment does not increase fatigue but will typically dilute the capital-labor ratio. See Feldstein for further discussion.

revenue product each week just equals their weekly earnings. Equation (11) says that the marginal benefit of increasing the length of the firm's workweek H_j should be set equal to the marginal cost, which is the number of workers employed times the increment to their earnings required to get them to work the extra time.

The second-order conditions, which are set out explicitly in my 1985 paper, are assumed to hold.

This treatment of the number of employees and the length of the shift as separate inputs allows for an explicit analysis of firm preferences for, say, layoffs instead of work sharing as a way of reducing labor input when demand falls. For example, by standard methods it can be shown that, under a reasonable additional assumption, the firm's level of employment and its workweek will depend positively on its output price, with the associated elasticities related to the shapes of both the production function and the earnings function. (See my 1985 paper, p. 13.) Thus, we may expect firms to react to depressed demand with both layoffs and work sharing, as indeed they did in the depression.[13] Similarly, it can be shown (under the same auxiliary assumption), that the firm's demand for workers and for hours per employee will decrease as reservation utility U^* rises (see my 1985 paper, pp. 13–14).

D. *Sectoral Equilibrium*

Determination of equilibrium employment and hours in the primary sector is now straightforward. It has been shown that the supply of employment increases with the level of utility U^* available in the primary sector, while the demand for employment can be expected to decrease with U^*. If there are n firms in the primary sector, the equilibrium level of reservation utility, call it U^{**},

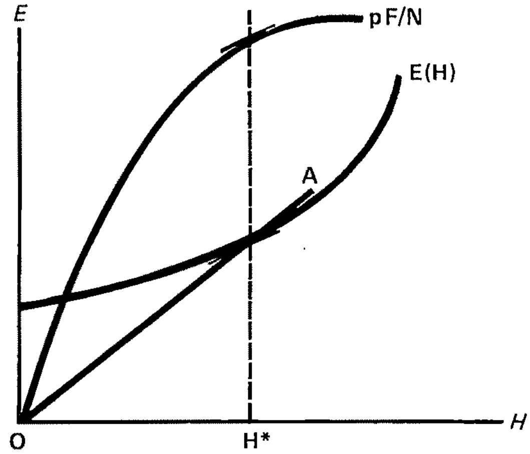


FIGURE 2. THE DETERMINATION OF H^*

satisfies

$$(12) \qquad N_t^s(U_t^{**}, \gamma_t) = \sum_{j=1}^{n} N_{jt}^d(U_t^{**}, p_t).$$

This level of utility is just enough to make the marginal worker indifferent between the secondary and primary sectors; inframarginal workers obtain a surplus in equilibrium.

Given U^{**}, p, and the N_j's, we may think heuristically of firms choosing hours of work H_j according to condition (11). (Of course, strictly speaking everything is determined simultaneously.) This is represented in Figure 2, which shows H_j^* being chosen at the level where the per capita total revenue curve (pF/N_j, written as a function of H_j) is parallel to the earnings function. Since the earnings function is an indifference curve, workers do not care what level of hours the firm chooses: Different firms in the industry (if they have different production functions) may well choose different workweeks in equilibrium. Indeed, since "wages" are simply average hourly earnings, and since earnings functions are not rays through the origin (recall the discontinuity at zero), firms using different workweeks may also be paying different wages in equilibrium. This result, which would be paradoxical in the traditional model, poses no problem here; workers comparing jobs look not at the wage but at

[13] This single-period rationalization of the layoff vs. work-sharing decision ignores an important dynamic element, i.e., the differential costs of adjusting work forces and shift lengths. This will be incorporated in the empirical model below.

the total utility (the combination of earnings and hours of work) available.

Although workers are indifferent about which point on the earnings function is selected by the firm, Figure 2 suggests an interesting observation. Note that a ray from the origin (OA) intersecting the earnings function at H_j^* in this case cuts through the earnings function from below. This illustrates the possibility that the average wage (E/H) may exceed the marginal wage (E_H) at the equilibrium level of hours. Thus, workers would happily work more hours at the *average* wage. That they are "constrained" not to do so is not the result of any market failure or disequilibrium, but of the difference between the marginal and average wage.

E. *Countercyclical Real Wages*

It is straightforward to generate countercyclical real (average) wages in this setup. Since the primary sector is by assumption cyclically sensitive, declining aggregate demand will cause the relative price of its output to fall. Without loss in generality, assume that the cost of living COL is unchanged while the output price p falls.

If N_j and H_j are normal inputs, firm (and industry) usage of both will fall as demand falls. For the moment, ignore the decline in N_j and consider only the effects of falling H_j:

Falling hours of work can be represented as a movement to the left on the earnings function (compare Figure 1). The reduced demand for hours will unambiguously reduce the *marginal wage*, E_H. However, the effect of falling hours on the *average* wage (and the average real wage, since COL is fixed) is ambiguous. The necessary condition for average wages to rise as hours fall is that the elasticity of earnings with respect to hours be less than one. This would be satisfied, for example, if U^* is not close to zero and the marginal disutility of labor does not increase quickly (i.e., the earnings function is close to linear, with a positive intercept). To anticipate, the empirical results do typically confirm that this elasticity is less than one.

The intuitive story underlying countercyclical real wages is as follows. A fall in industry demand causes employers to shorten the workweek. Workers will be benefited by the shorter hours of work, but will dislike the reduction in weekly earnings arising from short workweeks. The rate at which firms can reduce weekly earnings as the workweek falls depends on workers' preferences and reservation utilities. Especially at low levels of work and earnings, when consumption is highly valued relative to leisure, it may not be possible to cut weekly earnings as sharply as hours and still meet the reservation utility constraint. Thus the wage (i.e., hourly earnings) may rise even as labor demand and workweeks fall.

The iron and steel industry, as described by Carroll Daugherty et al. (1937) provides an illustration of these points. The workweek dropped extremely sharply during the 1930's in this industry. This occurred both because firms found it efficient to cut production by running certain operations only part time, and because firms' desires to maintain their work forces relatively intact led them to adopt "staggered" or "spread-work" schedules under which many workers worked only a few days a week (pp. 163–65). The problem posed by short workweeks for most workers was the obtaining of a basic sufficiency of income. It was estimated that in 1932–33 the weekly earnings of the average steelworker (not to mention the lowest paid) were less than half of that needed to reach a standard of "minimum health and decency" for a family of four (pp. 155–57). Moreover, "in most iron and steel communities there [were] few other opportunities for supplemental employment and income" (p. 167). Firms must have recognized that their ability to keep cutting total earnings as the workweek shortened was limited, since if workers could not attain a subsistence level in the mill town they would be forced to try elsewhere. Thus real hourly earnings in iron and steel rose, or fell relative slightly, as the workweek was cut.

The above discussion has emphasized the possibility that, *ceteris paribus*, a reduction in the workweek may tend to raise the real wage. This effect of falling hours of work will be offset to the degree that lower demand for industry output also results in lower

primary-sector employment N. Declining demand for employment, as well as any reduction in secondary-sector opportunities which result from the general downturn, will lower the equilibrium reservation utility level U^{**}. Lower U^{**} translates into a downward shift of the earnings function, which implies lower average wages for a given H. The net impact of the decline in the demand on average wages will depend on the relative strength of the various influences. In general, as in Lucas (1970), the cyclical behavior of the wage will be unrestricted.

An interesting implication of this analysis is that economies which rely more heavily on short workweeks (rather than employment reductions) as a way of reducing labor input are more likely to have countercyclical real wages. Using the same data as this paper, plus a matched data set for the postwar period, my paper with Powell verified that depression-era manufacturing industries did indeed exhibit both a greater relative reliance on variations in the workweek and greater countercyclicality in wages than did their postwar counterparts.

F. *Skilled and Unskilled Workers*

Until now I have assumed that workers are alike with respect to the production process. Empirically, it may well be the case that the changing skill mix over the cycle is of some importance. I consider this issue briefly now.

Suppose there are two types of workers, skilled and unskilled. Assuming that skilled and unskilled workers have systematically different opportunities in the secondary sector, or that they differ in number, they will have different supply functions. We can write the two supply functions in a given period, in the obvious analogy to (5), as $N_1^s(U_1^*, \gamma_1) = G_1(U_1^*/\gamma_1)$ and $N_2^s(U_2^*, \gamma_2) = G_2(U_2^*/\gamma_2)$, where the indexes 1 and 2 denote skilled and unskilled workers, respectively, and time and firm subscripts are suppressed. The corresponding earnings functions for the two groups are $E_i(H_i, U_i^*)$ where $U_i^* = \gamma_i G_i^{-1}(N_i)$ and $i = 1,2$. Normally, if $H_1 = H_2$, we expect to observe only N_i such that $E_1 > E_2$; that is, skilled workers earn more than unskilled.

On the demand side, assume that skilled and unskilled workers must be used in fixed coefficients, but that the ratio of skilled to unskilled falls as the length of the workweek (a proxy for the scale of production) expands. (See Rosen and references therein for evidence supporting this assumption.) Specifically, assume that if a firm is running the factory H hours per week, then a fraction $g_1(H)$ of its workers must be skilled and a fraction $g_2(H)$ unskilled, where $g_1 + g_2 = 1$ and $g_1(H)$ is decreasing in H. Then the firm's production function can still be written as in (8). Moreover, under the assumption that the skilled and unskilled must work the same number of hours, it is possible to write an average earnings function for the firm

$$\text{(13)} \quad \bar{E}(H, COL, U_1^*, U_2^*) = \sum_{i=1}^{2} g_i(H) E_i(H, COL, U_i^*),$$

where as before the U_i^* depend on total sectoral employment but are parametric to the firm.

The firm's optimization problem is not substantially complicated by the extension to skilled and unskilled workers (under the convenient assumptions that have been made). The firm finds optimal hours and total employment in precisely the same manner as in Part C above, except that the average earnings function $\bar{E}$ defined in (13) is used in place of the simple earnings function E. The division of employment into skilled and unskilled is then found by applying the ratios $g_1(H^*)$ and $g_2(H^*)$ to the optimal level of total firm employment.

The point of this digression is to highlight the effect of the changing skill mix on the properties of the empirically observed, average earnings function $\bar{E}$. Earnings functions defined for workers with identical utilities and productivities, for example, the E_1 and E_2, must be increasing, convex functions of hours, for $H > 0$. However, since the low-skilled and low-earnings fraction of the workforce varies procyclically, the *average* earnings function $\bar{E}$ will be flatter and have

a lower elasticity than *either* E_1 or E_2 taken separately. This has two implications. First, the empirically observed earnings function may not be convex in hours.[14] Second, as was shown above, a lower elasticity of the earnings function increases the probability of observing countercyclical real wages. Thus cyclicality of the skill mix may be an additional factor contributing to the solution of the wage puzzle.

G. *Implications for the Standard Approach*

The model set forth in the preceding parts of this section may seem outlandish to users of the standard labor market model (in which total worker-hours supplied and demanded are simply written as functions of the real wage). I believe, however, that the present model has a stronger prior claim than does the standard approach to being the correct way to model aggregate labor markets. The problem with the standard model is that, contrary to its major premise, workers are in fact typically not able to vary their labor supply continuously with respect to a parametric real wage; instead, they must choose among "packages" of total compensation, hours of work, and other job attributes offered by employers. The economic reason for the prevailing arrangement, as suggested above, is that there are usually economies or diseconomies of "bundling" of worker-hours. Supplying one hour of work each to eight different employers is not the same to a worker as supplying eight hours to a single employer. Similarly, employers are not indifferent between receiving one hour of work from eight different workers and receiving eight hours from one worker. As long as economies or diseconomies of bundling worker-hours exist, the standard model cannot be literally correct.

Inappropriate use of the standard model can lead to misconceptions. For example, the debate between supporters and defenders of the Lucas-Rapping intertemporal substitution approach has centered on the time series properties of the real wage (see, for example, Altonji and Ashenfelter). However, if the approach of the present paper is correct, rather than the standard model, then the behavior of the real wage is largely irrelevant to that debate.

Another example concerns the estimation of labor supply elasticities. Suppose workers have identical utility functions, given by

$$U = (E_t/COL_t) - \phi H_t. \tag{14}$$

That is, the marginal utility of consumption (equal to real earnings) and the marginal disutility of hours of work are constant. The labor supply elasticities of these workers (in the conventional sense) are infinite. But what will researchers using aggregate data and assuming the standard model find? Suppose that the model of the present paper actually applies, and that reservation utilities U_{it}^* are distributed in the population as assumed in Part B above. Then it is easy to show that the aggregate real wage w_t is given by

$$w_t = \left(\gamma_t G^{-1}(N_t)/H_t\right) + \phi. \tag{15}$$

Note that $\partial w/\partial N > 0$, $\partial w/\partial H < 0$. The observed relationship between worker-hours and the wage thus depends on whether N or H is more variable. Suppose, at one extreme, that the workweek is institutionally fixed: then the econometrician regressing total worker-hours against the wage will find a positive (although not infinite) elasticity, since all variation in worker-hours is attributable to changes in employment. At the other extreme, suppose that workweeks but not employment vary: then the estimated aggregate labor supply curve will be *backward bending*. The econometrician may well be concerned about the "instability" over time of his estimates, and their lack of relation to labor supply elasticities found in micro-level panel data. The problem, however, does not lie with data or identification

[14] It should be noted that that convexity of the earnings function is not necessary for the second-order conditions to hold. The earnings function estimated below is log linear, i.e., earnings is concave in hours; empirically, this seemed to work best.

TABLE 1—INDUSTRIES INCLUDED IN THE DATA SET

Industry (mnemonic)	Wage Earners[a]		Value-Added[b]	
	Thousands	% Total Mfg.	$Millions	% Total Mfg.
1. Iron and Steel (*IRON*)	419.6	5.02	1622.8	5.40
2. Automobiles (*AUTOS*)	226.1	2.70	1315.0	4.37
3. Meat Packing (*MEAT*)	122.5	1.46	460.5	1.53
4. Paper and Pulp (*PAPER*)	128.0	1.53	482.8	1.61
5. Boots and Shoes (*SHOES*)	205.6	2.46	450.9	1.50
6. Wool Textiles (*WOOL*)	179.6	2.15	414.8	1.38
7. Leather Tanning and Finishing (*LEATH*)	49.9	0.60	143.7	0.48
8. Lumber and Millwork[c] (*LUMBER*)	509.2	6.09	1088.5	3.62
Total	1840.5	22.01	5979.0	19.89

[a]Number of wage earners, and percentage of wage earners in all manufacturing employed in the industry, 1929; from Solomon Fabricant (1942, Appendix B).
[b]Millions of dollars of value-added, and percentage of all manufacturing value-added originating in the industry, 1929; from Fabricant (1940, Appendix C).
[c]Furniture is excluded.

problems, but with the use of the wrong model.

II. Empirical Implementation

This section begins with a brief description of the data used in this study. It then specifies an empirical model and reports the results of its estimation for each of the eight manufacturing industries. The estimated model is based closely on the analysis of the previous section but contains substantive additional elements as well.

A. *Data*

The data set constructed for this research includes, for each of eight manufacturing industries, *monthly* observations on the following variables: 1) production; 2) the (wholesale) price of output; 3) employment (of wage earners); 4) hours of work per week (per wage earner); and 5) average hourly earnings (of wage earners). The sample period runs from January 1923 to December 1939. The data from the 1920's were included so that the depression might be studied in a broader context, including a period of "normalcy." It is unfortunate that it was impossible to extend the sample even further back (see the Appendix).

The eight manufacturing industries covered, with measures of their relative importance, are listed in Table 1. The industries are diverse with respect to type of output (producers of durables, nondurables, and semidurables are represented), market structure, stage of development, geographical location, and the skill composition and demographics of the labor force. The choice of industries was not arbitrary; this was the largest set for which complete and reasonably consistent data series could be found. In particular, the desire to have series on weekly hours restricted me to industries regularly surveyed, beginning in the early 1920's, by the National Industrial Conference Board. (The Bureau of Labor Statistics, which surveyed many more industries, did not collect hours data before 1932.) Also, a number of candidate industries were eliminated by the requirement that the industrial produc-

tion series be based on a measure of the physical volume of output (for example, tons of iron), not inputed by using measures of inputs.[15]

Additional discussion of the data is contained in the Appendix. Also see my paper with Powell and my 1985 paper.

B. *Specification of the Supply Side*

The supply side in this model is summarized by the earnings function faced by each primary sector (manufacturing industry), as in equation (6). Average weekly nominal earnings received by wage earners in an industry have been shown to depend on four elements:

1) the length of the industry workweek H (now assumed to be the same for all firms);

2) industry employment N;

3) factors affecting workers' reservation utilities, γ;

4) the cost of living *COL*.

Hours and employment for each industry and the economywide cost of living are directly observed. The most difficult problem is to identify monthly determinants of workers' reservation utilities. Two factors which I expected to be important here were the level of government relief and the strength of the labor movement. As measures of these factors I constructed two (monthly) variables, *EMERGWORK* and *UNIONPOWER*. (A list of all variables used in estimation is given in Table 2.) *EMERGWORK* is the log of the number of "emergency workers" employed by the federal government, including all of the major work relief programs. *UNIONPOWER* attempts to capture the resurgence of the labor movement after the favorable legislation of the New Deal. This variable is set equal to zero until May 1935, the month the Wagner Act was passed. (The labor movement was extremely weak between the beginning of the sample period and 1935, as has been noted.) Starting with May 1935, *UNIONPOWER* is set equal to the cumulative number of man-days idled by strikes (in the economy as a whole).[16] The idea here is that strikes are an investment in the capital good of union credibility, which in turn affects the level of earnings workers are able to demand. (There is in fact a close correla-

TABLE 2—DEFINITIONS OF VARIABLES USED IN ESTIMATION

Variable	Definition
COL	Cost-of-living index
EARN	Nominal weekly earnings (per wage earner)
$\widetilde{EARN}$	Nominal weekly earnings, less intercept; see (17)
EMERGWORK	"Emergency workers" hired under the New Deal
EMP	Employment of wage-earners
EMPLF	*EMP* – *LABORFORCE*
HCOST	Marginal cost of *HRS*; see (24)
HRS	Weekly hours of work (per wage earner)
INTERCEPT	Intercept in earnings equation; see (17)
LABORFORCE	Aggregate labor force (Lebergott), interpolated
NRA	National Recovery Act dummy
P	Industry output price
PAY	Nominal weekly earnings deflated by product price
Q	Real production (not deseasonalized)
QSEAS	Purely seasonal component of real production
Q-QSEAS	Deseasonalized real production
t	Time
UNIONPOWER	Cumulative man-days idled by strikes; = 0, May 1935

Note: All variables except *COL*, *EMERGWORK*, *LABORFORCE*, *NRA*, *t*, and *UNIONPOWER* are defined separately for each of the eight industries. Other than *NRA*, *INTERCEPT*, *t*, and *UNIONPOWER*, variables are in logarithms.

[15]Aggregate industrial production indices are heavily contaminated by input-based measures of output. For this reason I obtained estimates only at the industry level, not for all manufacturing.

[16]Economywide rather than industry series are used primarily because of lack of data. Arguably, however, union successes in individual industries had "spillover" effects on industries not directly involved. (But see fn. 22 below.)

tion in this period between strike activity and the major new union recognitions and contracts that were achieved.)

The basic earnings functions that I estimated was of the form

$$(16)\quad \widetilde{EARN}_t = \alpha_0 + \alpha_H HRS_t + \alpha_E EMPLF_t + \alpha_W EMERGWORK_t + \alpha_U UNIONPOWER_t + \alpha_N NRA_t + COL_t + \alpha_t t$$

where

$$(17)\quad \widetilde{EARN}_t = \log(\textit{earnings} - \textit{cost of living} \times INTERCEPT)$$

and where the variables are as in Table 2. Equation (16) says that the log of nominal weekly earnings (less an intercept term, to be discussed in a moment) is a positive function of the log of hours worked per week, *HRS*; a positive function of the log of industry employment (normalized by national labor force), *EMPLF*; a positive function of workers' reservation utilities, as measured by *EMERGWORK* and *UNIONPOWER*; and is related one-for-one to the log of the current cost of living *COL* (i.e., there is no money illusion or imperfect information about price levels). Also included in the equation are a time trend (to capture secular influences on reservation utilities, such as demographics or wealth accumulation); and a dummy for the National Recovery Act period of September 1933–May 1935 (*NRA*), during which legislation affecting wages and hours may have had a direct impact on earnings.

The dependent variable of (16), which is defined in (17), is not simply the log of nominal weekly earnings, but the log of nominal weekly earnings less an expression which is constant when measured in real terms. This is in order to conform to a basic element of the theory, that there is a discontinuity in the earnings function at zero; that is, workers require some minimum pay "just to come to work." I did not expect to be able to estimate a value of the constant term *INTERCEPT* from the data, since sample values of hours worked are never very near zero and the earnings function is likely to be nonlinear in a relatively unrestricted way. Instead, for each industry I arbitrarily set *INTERCEPT* equal to the real value of six hours' pay at the rate paid in June 1929; that is, the "fixed cost of coming to work" was assumed to be equal to one hour's real pay for each day in the standard workweek. The exact value chosen for *INTERCEPT* was not at all crucial; I tried values from zero to twelve hours' pay without affecting the qualitative nature of the results.

Equation (16) was estimated, and was found to "work" fairly well empirically, in the sense that the estimated coefficients were of the right sign and were statistically significant. However, the estimated equations also had low Durbin-Watson statistics and did not perform particularly well in simulations. After some examination of the data, I recognized that the restriction in (16) that nominal earnings must be directly proportional to the current cost of living is not a good one. If this constraint were correct, it would imply that the high-frequency variation of earnings should be similar to that of the cost of living. In fact, the usual result that nominal labor compensation variables are "smoother" than price-level variables holds in these data.

To capture this smoothing effect, I assumed that nominal earnings respond only to the "permanent" component of cost-of-living changes, in the sense of John Muth (1981). That is, nominal earnings are proportional not to *COL* but to COL^*, where COL^* is defined by

$$(18)\quad COL^*_t = \lambda_P COL_t + (1-\lambda_P)COL^*_{t-1}.$$

Alternative interpretations of this assumption are that earnings are set each period on the basis of adaptive forecasts of the cost of living (compare Lucas and Rapping, 1969); or that costs of rapidly adjusting wage rates cause employers to attempt to smooth out the effects of cost-of-living changes (see Julio Rotemberg, 1982).

The final earnings function therefore was (16), with COL^*_t replacing COL_t. A Koyck

transformation of this equation, using (18), yields an observable model. Nonlinear estimation methods were used so that estimates of the original parameters α_H, α_E, α_W, α_U, α_N, α_t, and λ_P could be obtained. These results are discussed in conjunction with the demand-side results below.

C. *Specification of the Demand Side*

The primary constituent of the demand side of the model is the production function F. To obtain a specific functional form for F, I assumed that employment and hours of work aggregate as a generalized *CES*:

$$(19)\quad Q_t = B\left(\alpha e^{g_N t} N_t^{-\rho} + (1-\alpha) e^{g_H t} H_t^{-\rho}\right)^{(-k/\rho)},$$

where B, α, ρ, k, g_N, and g_H are parameters. This formulation allows for nonconstant returns to scale and factor-augmenting technical change. Explicit dependence of output on the capital stock and other nonlabor factors is suppressed because of lack of data; the hope is that these effects can be adequately represented, for the purposes of our short-run and medium-run analyses, by the exponential trend terms. (I experimented with quadratic as well as linear exponential trends in the estimation, without a significant effect on the results.) The expression (19) was chosen basically because it rationalizes simple log-linear relationships that have been shown to be empirically successful in other applications. Note however, that if the capital stock follows a time trend, then (19) is more general than some standard specifications, for example, the Cobb-Douglas form estimated by Feldstein.

With this specification, the first-order condition for employment (10) can be written as

$$(20)\quad n_t^* = \beta_{n0}^* + \beta_{nq}^* q_t - \beta_{ne}^*(e_t - p_t) + \beta_{nt}^* t,$$

where n^* is the log of employment; q is the log of output; $e-p$ is the log of weekly earnings divided by the output price; and the coefficients β^* depend on the production function parameters in a straightforward way.

The variable n_t^* may be thought of as the desired level of employment in period t; that is, it is the level of employment that exactly satisfies the first-order condition. We may suspect, however, that (20) will not be successful empirically, because costs of adjustment will prevent this relation from holding instantaneously (especially in monthly data, such as these). A possible response to this is to make the underlying model explicitly dynamic and solve the resulting maximum problem, as in Thomas Sargent's study (1978). Such an approach can become extremely complicated, however; and, as Sargent noted, it is not likely to reduce the need for auxiliary *ad hoc* assumptions. Here I follow the bulk of the previous work in simply assuming gradual adjustment of employment toward the desired level. That is, if n_t^* is the desired level of employment defined by the first-order condition, I assume that firms adjust actual employment n_t according to

$$(21)\quad n_t - n_{t-1} = \lambda_n(n_t^* - n_{t-1}),$$

where λ_n is the speed of adjustment. Given (21), a Koyck transformation of (20) gives an equation for actual employment of the form:

$$(22)\quad n_t = \beta_{n0} + \beta_{nq} q_t - \beta_{ne}(e_t - p_t) + \beta_{nt} t + \beta_{nn} n_{t-1},$$

where $\beta_{ni} = \lambda_n \beta_{ni}^*$, $i = 0,\ q, e, t$, and $\beta_{nn} = (1-\lambda_n)$.

Similarly, using the first-order condition (11) for desired hours of work, and making the reasonable assumption that there may also be costs to rapid adjustment of the length of the workweek,[17] we obtain

$$(23)\quad h_t = \beta_{h0} + \beta_{hq} q_t - \beta_{hc}(hcost_t) + \beta_{ht} t + \beta_{hh} h_{t-1},$$

[17]These include the costs of reorganizing production schedules and of inducing workers to rearrange their personal schedules.

where h is the log of average hours worked per week, *hcost* the log of the marginal cost of hours of work (as obtained from the earnings function; see below), and the coefficients are defined by the obvious analogies to the employment case, with λ_h the rate of adjustment of hours of work.[18]

Equations (22) and (23) may be viewed as representative-firm demand functions for employment and hours of work (length of the workweek). In both cases demand is positively associated with the level of production and negatively related to a cost variable. Except for the inclusion of the cost variables and the specification of separate equations for employment and hours, (22) and (23) are quite conventional short-run labor demand functions; see for example Frank Brechling (1965), Robert Ball and E. B. A. St. Cyr (1966), N. J. Ireland and D. J. Smyth (1967), and the survey in Ray Fair (1969), as well as Lucas and Rapping (1969).

In the construction of empirical versions of (22) and (23), several practical issues had to be addressed:

1) A basic question was the treatment of seasonality, which is fairly significant in these data. Fair has argued against deseasonalization in this context, on the grounds that the factors which explain cyclical movements in employment, etc., should also explain seasonal movements. There is also some danger that deseasonalization may introduce spurious relationships or obscure genuine ones. For these reasons the data were not deseasonalized prior to estimation, and seasonal dummies were not used in the equations. (Note that leaving in seasonal fluctuations causes an essentially spurious deterioration in fit.) I did, however, allow for the possibility that employment and hours demand might respond differently to the seasonal and nonseasonal components of production, as follows. For each industry I constructed a variable *QSEAS*, the "seasonal component of production," as the residual of the deseasonalization of industry output. I then allowed *QSEAS* and $Q - QSEAS$ (the seasonally adjusted component of production) to enter the demand for workers and hours equations with separate coefficients.

2) At an early stage of my analysis of this data set, I looked at the cross correlations, at various leads and lags, of the log-differences of output and the labor market variables (employment, hours, earnings). My concern was, given that the data are monthly and that the output and labor variables are from different sources, that there might be an alignment problem. This examination revealed little potential difficulty, except in the relation of the employment and output series: For a few industries, employment seemed more strongly related to output one month ahead than to current output. Given that the other labor series lined up with output, this seemed likely to reflect a genuine economic phenomenon, for example, hiring in advance of production, rather than a data alignment problem. In any case, for the employment demand equations, I allowed both current and one-month-ahead production to enter. (Since both seasonal and nonseasonal production were used, this gave a total of four output variables in these equations.) Actual one-month-ahead nonseasonal production was instrumented for rather than treated as exogenous in the estimation of the employment demand equations; thus its estimated coefficients may be interpreted as measuring the impact of one-month-ahead forecasts of output (rather than actual future output) on current employment. One-month-ahead seasonal production was taken to be exogenous, on the grounds that the recurring seasonal component should be perfectly forecasted.

The inclusion of one-month-ahead output did not appear necessary in the hours demand equation.

3) The marginal cost of extending the workweek one hour, *HCOST*, was defined for the empirical application by

$$(24)\quad HCOST = EMP + \widetilde{EARN} - HRS - P,$$

[18] For simplicity I have assumed that the adjustment of hours depends only on the difference between actual and desired hours, not on the difference between actual and desired employment (and similarly, for the adjustment of employment). Arguments made in M. Ishaq Nadiri and Rosen (1973) would favor the relaxation of this restriction.

where $\widetilde{EARN}$ is defined by (17) and P is the industry output price. This follows directly from (11) and the form of the earnings function (16). (*HCOST* is actually proportional to, not equal to, the marginal cost of increasing the workweek; the factor of proportionality will be absorbed into the estimated coefficient of *HCOST*.)

The marginal cost of adding a worker, *PAY*, is simply given by $EARN - P$; note that the intercept of the earnings function has no bearing on the construction of this cost variable.

4) Industry codes drawn up under the National Recovery Act imposed some direct constraints on firm employment and hours decisions, for example, through the work-sharing provisions. To allow for this, in the estimation the dummy variable *NRA* was added to both the employment and hours demand equations.

These considerations, in conjunction with equations (22) and (23), allow the specification of employment demand and hours demand equations that can be estimated for each industry. (For a list of the independent variables in these two equations, see the left-hand columns of Tables 3 and 4, and the variable definitions in Table 2.) The results of this estimation will be discussed in Part D below, following a digression on the identification problem.

C. *Identification*

The earnings equation and the two demand equations form a simultaneous system, which raises the standard estimation issues of identification and the availability of instruments. It was evident in this case that, as is often true, a strict application of the criteria for valid instruments would leave no instruments (except the constant and time), no identification, and no hope of proceeding further. In particular, it is difficult to come up with measured exogenous variables that are highly correlated with the fluctuations in the demands for industry outputs.[19]

After some consideration, I made the tactical decision to treat industry output as exogenous in estimation. Although the assumption of output exogeneity is not ideal, there are a few arguments in its favor (beyond the obvious one of necessity): First, there are many precedents (Lucas and Rapping, 1969, and virtually all papers in the traditional literature on the short-run demand for labor make this assumption). Second, and more important, treating output as exogenous seems likely to provide considerable identifying power at a relatively low cost in induced bias. Given the maintained presumption that fluctuations in aggregate labor demand, rather than labor supply, dominated prewar business cycles, the correlation of industry output with disturbances to the industry production function and earnings equations should be relatively small.

Besides output, other variables treated as exogenous included the cost of living and the government policy variables (*NRA*, *UNIONPOWER*, and *EMERGWORK*). Also, at some risk of bias in the presence of serial correlation, I treated lagged employment, workweeks, and earnings as predetermined variables. Given all of these assumptions, the three estimated equations are well-identified, both in the formal sense and in the sense that sharp estimates are obtained in the sample. However, it should not be forgotten that, given that some of the variables treated as exogenous are at best only approximately so, the results below should be interpreted with caution.

I did not attempt to enhance the degree of identification by imposing the cross-coefficient restrictions implied by the structural derivations of the various equations. My reason was that, because of the aggregation of the data, there is no serious reason to believe that such cross-equation restrictions will hold. The demand equations, for example, were derived for a hypothetical individual firm and will not literally apply (because logarithms do not add) to the industry-level data

[19]The money supply might seem to be a possible exception to this statement. I did experiment with this variable. However, its correlation with industry variables in monthly data is sufficiently low that it is of not much value as an instrument.

at hand. The strongest justifiable assumption, I believe, is that the qualitative magnitude and sign relationships survive the aggregation process; this is the assumption that underlies my interpretations of the results.

Note that, given that the previous assumptions make the cross-coefficient restrictions inessential to identification, failure to impose them at worst may cause a small loss in efficiency. This is not a serious issue, given the size of the data set.

D. *Estimation Results*

We proceed now to the estimates. The results of estimating the demand equations are in Tables 3 and 4; the earnings function results are in Table 5. Two-stage least squares (*2SLS*) was used to correct for simultaneity bias (for the earnings equations, nonlinear *2SLS* was used). Instruments and variables treated as endogenous are given in the Appendix. Each equation was estimated separately for each industry.

The employment and hours demand equations are modest extensions of the conventional formulation and should be uncontroversial; they will be discussed first. (See Tables 3 and 4.) The estimates suggest the following.

First, there appear to be significant costs of adjustment (or some other source of inertia) for both employment and weekly hours of work; that is, the lagged value of the dependent variables shows up as highly significant in every case. We would expect employment to be more inertial than hours of work, and this is confirmed by the estimates in every industry except automobiles. The rates of adjustment implied by the estimates are rather rapid: on average, the industries are able to eliminate about one-quarter of the gap between actual and desired employment each month, and nearly one-half of the gap between actual and desired hours.

For both employment and hours demand, the cost variables (*PAY* and *HCOST*) enter with the expected negative signs for each industry (except for one case, in which the coefficient of *PAY* is effectively zero). However, the statistical significance of the cost variable in the employment demand equation is low for some industries. The low significance of *PAY* is possibly due to the use of monthly data, which may obscure the presumably slow substitution between workers and other factors of production. The effect of the cost variable in the hours demand equation, in contrast, tends to be large and is in each case highly statistically significant. The lower inertia of hours of work and its greater sensitivity to short-run cost changes suggest that workweeks will lead employment in cyclical downturns; this conforms to the findings of Geoffrey Moore (1955), Gerhard Bry (1959), and myself and Powell.

The effect of production on input demand is broken down, for employment, into the effects of current "seasonal" production, current nonseasonal production, and seasonal and nonseasonal production one month in the future. Not too much systematic emerges from this breakdown; in particular, employment in some industries seems to depend most strongly on current output, while other industries hire "one-month ahead." However, although a few negative signs are scattered through the estimated effects of components of production on employment, the total estimated effects of output on employment are, as expected, positive and strongly significant. (See the last row in Table 3.) In conjunction with the estimated speeds of adjustment, the estimated output effects confirm in some cases, although not all, the familiar finding of short-run increasing returns to labor.[20]

In the hours equation, it was necessary to consider only current output effects. As can be seen from Table 4, both the seasonal and nonseasonal components of production, and

[20] The "total output effect" coefficients in the last row of Table 3 (and Table 4) actually double count the effect of an output increase on employment (or hours), since they represent the effect of a simultaneous increase in adjusted output and the (multiplicative) seasonal adjustment factor. Short-run increasing returns exists when the sum of the coefficients on either seasonal or nonseasonal output alone, divided by one minus the coefficient on the lagged endogenous variable, is less than one.

TABLE 3—INDUSTRY DEMANDS FOR WORKERS
(Dependent variable: EMP_t)

Independent Variables	IRON	AUTOS	MEAT	PAPER	SHOES	WOOL	LEATH	LUMBER
EMP_{t-1}	.740	.610	.916	.881	.720	.578	.771	.684
	(19.6)	(11.3)	(16.7)	(42.5)	(13.2)	(11.8)	(19.2)	(16.1)
PAY_t	−.135	−.134	−.094	−.046	−.202	−.229	.008	−.204
	(−3.70)	(−1.04)	(−1.98)	(−1.59)	(−1.87)	(−2.82)	(0.16)	(−2.91)
$Q_{t+1} - QSEAS_{t+1}$	.123	.450	.918	.179	−.033	.002	−.046	.216
	(1.69)	(5.39)	(1.87)	(1.27)	(−0.32)	(0.02)	(−0.26)	(1.39)
$QSEAS_{t+1}$	.102	.343	.124	.039	.220	.286	.147	.279
	(1.94)	(4.94)	(1.90)	(0.98)	(8.12)	(2.61)	(2.36)	(3.50)
$Q_t - QSEAS_t$	.058	−.125	−.532	−.050	.240	−.008	.236	.071
	(1.94)	(−2.06)	(−.137)	(−0.37)	(1.86)	(−0.11)	(1.37)	(0.42)
$QSEAS_t$	.081	−.160	−.016	.082	−.030	.393	.160	.172
	(1.47)	(−2.35)	(−0.23)	(2.01)	(−1.03)	(5.10)	(2.55)	(1.73)
NRA_t	.010	.035	.026	.019	0.18	−.002	.015	−.024
	(1.04)	(1.48)	(1.40)	(2.72)	(1.85)	(−0.16)	(2.38)	(−1.04)
Durbin-Watson	1.46	1.79	1.81	2.11	1.90	1.91	1.65	1.55
Sum of Output	.364	.508	.494	.250	.397	.673	.497	.737
Coefficients	(6.71)	(6.69)	(3.68)	(4.76)	(4.74)	(8.06)	(6.69)	(7.34)

Notes: Sample: January 1923–December 1939; estimation was by *2SLS*. See Table 2 for variable definitions; instruments are given in the Appendix. Estimates of the constant and the trend term are not reported. The *t*-statistics are shown in parentheses.

TABLE 4—INDUSTRY DEMANDS FOR HOURS OF WORK
(Dependent variable: HRS_t)

Independent Variables	IRON	AUTOS	MEAT	PAPER	SHOES	WOOL	LEATH	LUMBER
HRS_{t-1}	.560	.761	.521	.612	.397	.512	.562	.381
	(14.0)	(11.0)	(10.6)	(14.5)	(5.58)	(10.1)	(10.0)	(5.83)
$HCOST_t$	−.312	−.162	−.116	−.056	−.509	−.334	−.217	−.159
	(−7.76)	(−2.85)	(−6.88)	(−2.51)	(−6.42)	(−7.82)	(−5.22)	(−3.70)
$Q_t - QSEAS_t$	.323	.131	.196	.220	.474	.242	.130	.254
	(12.4)	(4.01)	(7.11)	(8.70)	(7.50)	(9.17)	(5.05)	(7.44)
$QSEAS_t$	.231	.062	.111	.162	.160	.348	.263	.357
	(3.99)	(1.96)	(5.01)	(3.49)	(3.29)	(5.31)	(3.63)	(5.70)
NRA_t	−.041	−.010	−.016	−.026	.047	−.042	−.007	−.062
	(−3.40)	(−0.53)	(−2.34)	(−4.86)	(3.06)	(−3.95)	(−0.80)	(−4.61)
Durbin-Watson	1.99	1.57	1.73	1.80	1.47	1.44	1.49	1.54
Sum of Output	.553	.193	.307	.382	.634	.590	.394	.611
Coefficients	(8.17)	(3.42)	(7.54)	(7.06)	(6.42)	(7.65)	(4.94)	(6.89)

Notes: See Table 3.

of course their sum, have a strongly significant, positive effect on hours of work. Short-run increasing returns to this factor appear to exist for all industries.

The final estimated parameters show the effects of the NRA codes on industry demands for labor inputs. The results imply that, for the most part, the NRA tended to increase employment and reduce hours. This is consistent with one of the legislation's explicit goals, which was to increase employment through "work-sharing." It also helps to explain the persistence of part-time work during the post-1933 recovery.

The residual serial correlation in the two demand equations appears to be relatively low, although it must be remembered that the Durbin-Watson statistic will be biased by the presence of the lagged dependent variables. (Calculated values of Durbin's h-statistic, which corrects for the lagged dependent variable problem, implied that the hypothesis of *no* serial correlation could be rejected for each equation; however, this statistic gives no information about the extent of serial correlation.) Reestimation of the demand equation using Fair's method gave qualitatively similar results. I also estimated the two demand equations for each industry jointly, so that contemporaneous correlation of residuals could be accounted for; this led to virtually identical results. For computational reasons, I did not attempt to estimate any equation for all industries simultaneously.

In tests for stability across the 1920's and 1930's subsamples, five of the eight employment equations and four of the eight hours equations failed at the .05 significance level. This is not really surprising, given the stark differences in the economic environments of the two periods. When the demand equations are estimated for the subsamples separately, however, they do not look grossly different. In particular, estimates for the 1930's subsample, as well as for 1923–29 and 1923–33, have the right signs and look very much like the whole-sample results.

Overall, the estimated labor demand equations seem reasonably successful, certainly of sufficient quality to use in simulation exercises. They also lend some support to the treatment of employment and hours as "separate" factors of production.

Estimates of the earnings equations, which make up the "supply side" of the model, are given in Table 5. The estimated parameters are those defined in equations (16) and (18): The most important of these are α_E and α_H, which capture the sensitivity of earnings to employment (normalized by the labor force) and to hours of work, and λ_P, which measures the speed of adjustment to cost-of-living changes. I have reported separate estimates for 1923–33 (which avoids the effects of New Deal legislation; i.e., $\alpha_N = \alpha_U = \alpha_W = 0$) and for the whole period.

If we look first at the results for 1923–33, we find that overall the results conform closely to the predictions of the theory. First, for a given level of weekly hours there is typically a strong positive relationship between earnings and employment. This is interpretable as a supply relationship; that is, to induce more workers to enter an industry, firms must increase the utility value of the earnings-hours packages they offer. Second, the elasticity of earnings with respect to hours of work is highly significant, positive, and typically less than one;[21] as argued above, finding this elasticity to be less than one is consistent with countercyclicality of real wages. Finally, nominal earnings adjust only partially to current changes in the cost of living; for the 1923–33 sample, the average rate of adjustment is about 17 percent per month. Although this is a significant amount of "stickiness," it is much less than is usually assumed by Keynesians.

One industry that looks somewhat different from the others is automobiles. For both sample periods, the measured sensitivity of earnings to employment is low, while the sensitivity of earnings to hours is the highest of any industry. The earnings function for the automobile industry was even more striking when it was reestimated without an imposed intercept (i.e., *EARN* rather than $\widetilde{EARN}$ was used as the dependent variable). In that case, for both sample periods, the elasticity of earnings with respect to employment was almost exactly zero and the elasticity of earnings with respect to hours was almost exactly one. This result (which was quite different from what was obtained for the other industries) would be consistent with an industry policy of setting a flat wage rate, which is not changed even when the workweek changes, and of rationing the

[21]Actually, the estimated coefficient α_H measures the elasticity of earnings less the intercept, not earnings itself, to hours. The elasticity of earnings to hours is strictly less than α_H and is less than one for each industry except automobiles.

TABLE 5—EARNINGS FUNCTIONS

Estimated Parameter	*IRON*	*AUTOS*	*MEAT*	*PAPER*	*SHOES*	*WOOL*	*LEATH*	*LUMBER*
1. Sample: January 1923–June 1933								
α_E	.352	.048	.202	.496	1.111	.285	.267	.364
	(4.13)	(0.98)	(2.71)	(2.95)	(5.76)	(4.95)	(3.07)	(3.33)
α_H	.951	1.172	.648	.869	.784	.737	1.010	.817
	(11.8)	(22.1)	(8.79)	(7.43)	(7.53)	(9.22)	(17.8)	(4.16)
λ_P	.127	.173	.188	.078	.204	.175	.145	.320
	(3.19)	(2.99)	(3.88)	(2.57)	(3.30)	(3.97)	(3.40)	(4.52)
Durbin-Watson	1.99	1.82	1.96	2.16	2.09	1.97	2.25	2.20
2. Sample: January 1923–December 1939								
α_E	.320	.004	.118	.419	.317	.326	.252	.369
	(4.08)	(0.14)	(1.85)	(2.84)	(2.46)	(5.76)	(3.90)	(4.28)
α_H	1.030	1.203	.713	.913	.983	.697	.966	.698
	(18.9)	(27.0)	(8.52)	(11.8)	(12.2)	(9.13)	(18.1)	(4.76)
α_N	.029	.020	.009	.008	.053	.057	.040	.036
	(1.47)	(1.25)	(0.53)	(0.64)	(2.38)	(3.11)	(3.73)	(0.98)
α_U	.229	.183	.210	.137	−.250	.117	.097	.030
	(2.42)	(1.61)	(4.00)	(2.23)	(−1.60)	(1.63)	(2.32)	(0.24)
α_W	−.069	−.131	.163	.007	.008	−.035	.045	.020
	(−0.88)	(−1.94)	(2.64)	(0.17)	(0.08)	(−0.47)	(1.03)	(0.17)
λ_P	.129	.073	.204	.095	.068	.163	.157	.217
	(3.84)	(2.22)	(4.81)	(3.27)	(1.97)	(4.45)	(4.50)	(4.60)
Durbin-Watson	1.83	1.84	2.03	2.20	1.95	1.97	2.10	2.32

Note: Estimation was by *NL2SLS*. See text for parameter definitions; instruments are given in the Appendix. Estimates of the constant and trend term are not reported. The estimates of α_U and α_W are multiplied by 10^5 and 10^4, respectively, for legibility. The *t*-statistics are shown in parentheses.

available jobs among applicants. It is worth noting in this connection that Henry Ford was a prominent maverick of this time in wage and employment policies; a fixed, high wage plus job rationing might not be a bad description of his announced strategy for improving worker motivation.

The estimates of the basic parameters for the whole sample (the bottom half of Table 5) are fairly similar to those for 1923–33, although the rate of adjustment of earnings to prices is estimated to be under .1 in three cases (instead of in just one case for 1923–33). The major difference is that the equation for the 1923–39 period also incorporates estimates of the effects of the New Deal on earnings. Briefly, the estimates show, first, that the NRA codes had relatively small but positive effects on weekly earnings. Second, the expansion of union power after the Wagner Act appears to have had a strong positive impact on earnings, raising weekly earnings by about 10 percent or more in six of the industries. (In lumber, the effect of unionization appears to have been positive but negligible; in boots and shoes, workers suffered significant pay *cuts* during the late New Deal.)[22] Finally, government employment programs appear to have had little systematic effect on the earnings of those privately employed in manufacturing.[23]

[22] Horace Davis notes: "Another significant point [regarding the decline in shoe industry wages], as bearing on the year 1937, was the checking of the unionization drive in shoes at the very time when unionism was getting established in several other manufacturing industries for the first time" (1940, p. 98). The unusual decline in shoe industry wages after 1937 probably also accounts for the very different estimates of α_E in the 1923–33 and the 1923–39 samples.

[23] Henning Bohn suggested that agricultural earnings, an additional measure of workers' alternative opportunities, might belong in the industry earnings functions; so I tried this. Monthly agricultural wage rates (nominal, without board) are reported for each quarter in the sample in R. A. Sayre (1940); I interpolated this series

For both the short and long samples, diagnostic checks did not seem to indicate important amounts of serial correlation. Because of this, and because the nonlinear version of Fair's method imposes some computational costs, I did not make any serial correlation correction.

I performed some additional diagnostic analyses of the estimated equations. Of these, the most interesting were within-sample simulations of the complete model. Space does not permit a full reporting here of these experiments (see my 1985 paper for more detail); but I will note that, in dynamic simulations of both 1930–33 and the New Deal era, the model did quite well overall (according to a number of criteria) in tracking the major variables. In particular, the model simulations did a creditable job of tracking the real wage in each of the eight industries, the tendency of real wages to rise even as output and employment fall being clearly exhibited.

Three factors contributed to the model's ability to simulate countercyclicality in wages: namely, the tendency in this model for wages to be countercyclical when workweeks are cyclically sensitive, as was discussed above; the assumed inertia in nominal wages (important in 1930–33); and unionization effects (important after 1935). In order to obtain an idea of the relative importance of nominal wage inertia in the determination of real wages in the critical 1930–33 period, I conducted the following experiment. I ran the simulations of 1930–33 again, this time assuming perfect adjustment of nominal earnings to the cost of living ($\lambda_p = 1$). All other coefficients were unchanged. I found that, first, although a rising real wage was still predicted by the simulations, the ability of the simulations to track the actual real wage deteriorated significantly for most of the industries. In several cases the root mean square error of simulation increased by one-half or more; also, the maximum real wage over the period predicted by the simulations tended to be quite a bit lower than what was actually attained. Thus, although not the whole story, a degree of nominal wage inertia seems to be an essential element in the explanation of real wage behavior in the early depression. There was also, however, a rather surprising second finding from these simulations: the assumption of perfect wage adjustment to the cost of living had virtually no effect on the ability of the model to track employment and hours. Indeed, on average, fits improved slightly. Thus, despite the importance of lagged adjustment for explaining observed real wage behavior in this period, this phenomenon may not have had great allocative significance.

Perhaps as interesting as the successes of the model in simulations were its occasional failures. For example, the model did not predict a strong attempt in 1932 by the steel and automobile industries to preserve their workforces through pronounced work-sharing strategies (i.e., a sharp cut in hours of work coupled with significantly increased employment). Another problem was that the model simulations tended to understate somewhat the degree of nominal inertia of wages during the first six to nine months of 1931. The first problem probably reflects the oligopolistic nature of the two particular industries, and their resulting ability to deviate from competitive behavior in the short run; the second difficulty no doubt results from the assumption that the sensitivity of wages to cost-of-living changes was the same in all periods, rather than being dependent on recent price behavior. (Since the 1920's were a period of stable prices, presumably the sensitivity of wages to prices was in fact less in the early 1930's than later on.) Significantly, however, these deviations from the projected paths were in each case quite transitory, with the simulated variables returning to their predicted paths in a year or less. Thus, al-

and divided by the cost of living to obtain a monthly series on real agricultural earnings. Reestimated earnings functions including this variable looked quite similar to those reported in Table 5. The estimated coefficient of agricultural earnings was typically found to be positive, as predicted by the theory, but of only moderate magnitude and statistical significance. An exception was the lumber industry, for which agricultural wages appear to have had an important influence on earnings.

though these failures suggests ways of improving the model, they do not appear to present fundamental difficulties.

III. A Dynamic Labor Supply Equation

A possible objection to the supply side of the model developed and estimated in this paper is that it is rather static in nature. I have made strong assumptions (that workers cannot borrow or lend, and that they have intertemporally separable utility functions) in order to avoid consideration of the intertemporal substitution of leisure and consumption. In addition, the implicit assumption that there are no mobility costs to moving between the secondary and primary sectors implies that workers need consider only current returns (and not long-run returns) when deciding whether to change sectors. Only the partial adjustment of nominal earnings to cost-of-living changes (in the estimated earnings functions) induces a modest dynamic element.

Although developing a more explicitly dynamic representation of this paper's model of labor supply is not particularly difficult conceptually, there are some substantial problems of empirical implementation. Rather than tackle those here, I propose to do something more limited: I will try to show that one of the more empirically successful models of depression-era labor supply, the intertemporal substitution model of Darby, can be reinterpreted as a dynamic version of the supply model in this paper. Estimates of a Darby-type model on these data will then be presented. The reasonableness of these estimates, it will be argued, constitutes evidence that the present paper's model of labor supply could survive the transition to a more dynamic specification.

Darby's model of labor supply is an extension of the basic Lucas-Rapping (1969) formulation. Lucas and Rapping argued, it will be recalled, that labor supply (i.e., worker-hours, normalized by the number of households) should depend 1) *positively* on the current returns to working; 2) *negatively* on the long-run, or "normal" returns to working; and 3) *negatively* on the ratio of the normal to the current price level.

The reasoning should be familiar. 1) High current returns to work increase labor effort by improving the rate of exchange between work and consumption. 2) High long-run returns to work depress current labor supply by making it more profitable to substitute present for future leisure. 3) Finally, assuming that nominal interest rates do not adjust fully to inflation, an increase in the ratio of the normal to the current price level lowers labor supply by leading workers to anticipate lower real rates of interest.

An important issue in this context, and one that I do not believe has been adequately addressed by the intertemporal substitution literature, is how to measure the returns from working. Lucas and Rapping, and most other authors, have assumed that the real wage is a good proxy for these returns. However, as was discussed in the introduction, this assumption makes it hard to explain labor supply behavior in the 1930's.[24] One of a number of contributions made by Darby was the substitution of full-time-equivalent earnings (*FTE*)[25] for the wage as the measure of the returns to work. Darby showed that using earnings instead of wages significantly improved the capacity of the model to fit the 1930's.

What is the rationale for using earnings rather than wages as a measure of the returns to work? Darby's argument was that, because the NRA codes required shorter workweeks, actual hours of work either were underreported by firms (leading to an upward bias in the measurement of hourly wages) or, possibly, were rationed. For these reasons he expected average earnings per *FTE* employee to "more accurately reflect the development of wages in the 1930s" (p. 10) than the official wage series.

[24] It may be noted also that estimation of a Lucas-Rapping-type model using these data (equations (25)–(27) below with the real wage in place of real earnings) yielded a number of wrong signs and a generally poorer fit than the Darby real earnings version.

[25] The *FTE* earnings variable used by Darby is essentially identical to actual average earnings for most industries, including manufacturing. That is, the variable reflects actual rather than normal workweeks. See the *Survey of Current Business*, June 1945, pp. 17–18.

A problem with Darby's argument is that the NRA codes were in effect for less than two years, but the substitution of earnings for wages seems to be empirically preferable for the entire prewar period. (See Darby's paper, and the results below.) An alternative explanation for the superiority of the earnings variable follows from the analysis of the present research. It has been suggested here that, in an environment where hours of work are not constant, the correct measure of the returns to working is neither wages nor earnings but the total utility of the earnings-hours package offered by the job, perhaps measured relative to the utility of remaining in the secondary sector. An obvious problem, however, is that this utility is not observable to the econometrician; thus we might ask which, if any, of the observables is likely to be correlated with the total utility of a job. The wage is not a good choice; as has been shown at length, wages and the utility of a job can easily move in opposite directions. However, in the case where fluctuations in employment are due primarily to variations in demand rather than supply (the probable situation in the 1930's), the utility from holding a job and earnings will be highly correlated. This is straightforward to show. Increased demand in the primary sector, which increases the equilibrium utility of workers, will also typically both move the equilibrium earnings function upward and increase hours of work. Thus increased primary-sector demand will also increase earnings. The explanation for the superiority of Darby's specification, therefore, is simply that earnings are a good proxy for the total utility of holding a job, and wages are not.

These considerations suggest that estimating a model in the spirit of Darby on the present data set may be a valuable exercise. I specify an empirical model as

$$(25)\quad EMP_t \times HRS_t - LABORFORCE_t = \beta_0 + \beta_1(EARN_t - COL_t) + \beta_2(EARN_t - COL_t)^* + \beta_3(COL_t^* - COL_t) + \alpha_N NRA_t + \alpha_U UNIONPOWER_t + \alpha_W EMERGWORK_t + \alpha_t t$$

$$(26)\quad (EARN_t - COL_t)^* = \lambda_P(EARN_t - COL_t) + (1-\lambda_P)(EARN_{t-1} - COL_{t-1})^*$$

$$(27)\quad COL_t^* = \lambda_P COL_t + (1-\lambda_P)COL_{t-1}^*$$

where an asterisk denotes the "permanent" or long-run component of a variable. (Variables definitions are given in Table 2.)

Equation (25) is a labor supply equation, of the general form first written down by Lucas and Rapping (1969). The dependent variable is total worker-hours supplied to an industry, normalized by Lebergott's aggregate labor force estimates. (Lebergott's annual data were linearly interpolated to obtain a monthly series.) Equation (25) follows the discussion above in specifying that the supply of worker-hours to an industry depends differentially on the current and long-run returns to working (where the returns to work are measured by real weekly earnings), as well as on the ratio of the long-run to current cost of living. The use of earnings to measure the returns to work reflects Darby's innovation. By the logic of the intertemporal substitution model, the expected signs of the coefficients are $\beta_1 > 0$, $\beta_2 < 0$, and $\beta_3 < 0$.

The labor supply equation (25) also contains terms reflecting New Deal government actions. The expected signs of the coefficients are: for α_N, ambiguous (since the NRA codes increased employment but reduced hours); for α_U, negative (since unionization should restrict labor supply below competitive levels; and for α_W, negative (since increased public works programs should reduce the supply of labor to industry). A time trend is also included.

Equations (26) and (27) follow Lucas and Rapping in assuming that the permanent components of returns and the cost of living are updated adaptively, with the same "rate of learning" applying in both cases. Constants and trends are excluded from (26) and (27); if included they would be absorbed into the constant and trend of the estimated equation, with no effect on the important estimated parameters.

Using (26) and (27), the labor supply equation (25) can be transformed so that

TABLE 6—DYNAMIC LABOR SUPPLY EQUATION

Estimated Parameter	*IRON*	*AUTOS*	*MEAT*	*PAPER*	*SHOES*	*WOOL*	*LEATH*	*LUMBER*
Sample: January 1923–June 1933								
β_1	1.78	1.63	1.77	1.45	1.04	2.10	1.21	1.80
	(8.25)	(11.1)	(6.98)	(8.51)	(12.1)	(10.1)	(11.1)	(3.24)
β_2	0.03	2.29	2.99	0.30	0.29	0.42	2.92	0.26
	(0.08)	(3.43)	(0.97)	(0.41)	(0.83)	(0.67)	(2.31)	(0.58)
β_3	−0.95	−0.66	−2.55	−1.51	−0.74	−2.69	−1.24	−3.12
	(−0.74)	(−0.46)	(−3.32)	(−3.55)	(−1.24)	(−2.15)	(−2.47)	(−1.53)
λ_P	.137	.148	.093	.071	.184	.199	.095	.351
	(2.57)	(3.49)	(1.59)	(2.30)	(3.17)	(3.26)	(2.72)	(4.24)
Durbin-Watson	2.02	1.85	1.81	2.06	2.04	2.23	1.86	2.08
2. Sample: January 1923–December 1939								
β_1	1.43	1.99	1.86	1.38	1.38	2.24	1.31	1.59
	(16.3)	(10.5)	(7.54)	(12.5)	(15.2)	(14.8)	(12.9)	(8.28)
β_2	0.52	2.61	4.11	0.79	0.10	0.67	3.01	0.52
	(2.05)	(3.34)	(1.62)	(1.71)	(0.23)	(1.25)	(3.35)	(1.46)
β_3	−1.93	−0.32	−3.11	−1.64	0.07	−2.01	−1.53	−4.21
	(−2.74)	(−0.21)	(−4.39)	(−5.97)	(0.12)	(−2.35)	(−4.17)	(−3.99)
α_N	−.021	.085	−.003	.011	.020	.032	−.009	−.046
	(−0.49)	(0.98)	(−0.06)	(0.65)	(0.62)	(0.66)	(−0.42)	(−0.87)
α_U	−.035	−.102	−.088	−.030	−.051	−.009	−.027	.023
	(−1.56)	(−2.57)	(−2.04)	(−1.92)	(−3.24)	(−0.43)	(−1.98)	(1.00)
α_W	.142	.098	−.149	.019	−.145	.157	.065	−.211
	(0.86)	(0.31)	(−0.87)	(0.28)	(−1.10)	(0.89)	(0.76)	(−1.03)
λ_P	.128	.170	.094	.081	.136	.165	.107	.174
	(3.90)	(3.89)	(2.49)	(3.68)	(2.76)	(4.03)	(3.99)	(4.52)
Durbin-Watson	1.85	1.79	1.82	2.13	2.21	2.07	1.93	2.20

Notes: See Table 5: only exception is the estimates of α_U and α_W are multiplied by 10^4 for legibility.

only observable variables appear (see Lucas and Rapping). The use of a nonlinear estimation procedure permits the recovery of the original parameters of (25) to (27).

The results of estimating the system (25)–(27) are reported in Table 6. The estimation method was (nonlinear) two-stage least squares, used to correct for simultaneity bias. (Instruments are listed in the Appendix.) The sample was January 1923–December 1939; estimates obtained for the sample ending before the New Deal, which set $\alpha_U = \alpha_W = \alpha_N = 0$, are also reported. The Durbin-Watson statistics are for the Koyck-transformed equations.

The most important result in Table 6 is that the estimate of β_1, which measures the elasticity of worker-hours supplied to earnings, is positive and highly significant in every case. There is also a remarkable uniformity across industries and sample periods of the magnitude of this estimated parameter. This is consistent with the idea that (25) is a true supply curve in which earnings are acting as a proxy for the total utility from working.[26]

The estimates of β_2 are also all positive, although magnitudes and statistical significance vary. The finding that β_2 is positive, that is, that higher long-run returns to work increase labor supply, is the opposite of the prediction of the intertemporal substitution model. An explanation of this finding is available, if we are willing to reinterpret (25). Recall that these estimates have been obtained from industry-level, not aggregate, data. At the level of the industry, labor supply depends not only on the decisions of workers

[26]The positive and significant estimates of β_1, it should be noted, did not simply reflect the fact that weekly hours is a constituent of both the dependent variable and weekly earnings. Reestimates using employment as the dependent variable instead of worker-hours also yielded positive and highly significant values for $\hat{\beta}_1$.

already "in" the sector (for example, already living in the mill town), but also on the number of workers that can be drawn from the rest of the economy. If there are mobility costs to switching sectors, higher long-run returns in an industry will *increase* the industry's labor supply, by making it more worthwhile for workers to incur the fixed costs of entering the sector. Thus it might be argued that long-run earnings belong in (25) because of their relevance to worker mobility decisions, not for any reason of intertemporal substitution.

This alternative interpretation of (25) is an attractive one, and not simply because it rationalizes $\beta_2 > 0$. A drawback of the intertemporal substitution hypothesis as a model of 1930's labor supply is that it assumes perfect capital markets. This assumption appears at variance with the tremendous disarray of the financial sector in the depression, and the resulting large difference between lending and borrowing rates for consumers.[27] In contrast, the mobility-cost interpretation of (25) does not require perfect capital markets; indeed, under this interpretation (25) is consistent with the Section I model of this paper, with its no-borrowing, no-lending assumption.

With respect to the effects of the New Deal, Table 6 finds the same result as the estimated earnings function in Table 5; namely, that the legislation-supported unionization drive was the most important New Deal change in labor markets. In contrast to the NRA codes and government work programs, which had little systematic impact, unionization appears to have had a strong effect in a number of industries.

Overall, the Darby-type specification seems to work well in these data. If the interpretation of this specification that I have given is accepted, this bodes well for the development of a more explicitly dynamic version of this paper's model of labor supply.

[27]See my 1983 article. The failure of the perfect capital markets assumption may explain the difficulty the intertemporal substitution model has in explaining the path of consumption in the 1930's (Altonji).

IV. Conclusion

This paper has employed monthly, industry-level data in a study of Great Depression labor markets. The framework of analysis was a model in which, as in Lucas (1970), both firms and workers are concerned with the distinction between the number employed and the number of hours each worker works. In the context of the depression, this distinction appears to be an important one; and, in conjunction with additional empirical elements, this model does a rather good job of explaining the behavior of the key time-series. This raises the possibility that the decomposition of aggregate labor supply into participation rates and hours per worker may be important for understanding other macroeconomic episodes as well.

A limitation of this analysis is its partial equilibrium nature: output is treated as exogenous. A really satisfactory analysis of the 1930's would have to consider labor markets, product markets, and financial markets in a simultaneous general equilibrium. This should be pursued in future research.

Appendix

The sources of the data used in this study are as follows.

1) Earnings, hours, and employment data are from M. Ada Beney (1936) and R. A. Sayre (1940). These data are the result of an extensive monthly survey conducted by the National Industrial Conference Board from 1920 until 1947.

All of the industries in the sample paid at least part of their workforce by piece rates (see the *Monthly Labor Review*, September 1935, pp. 697–700). No correction was made for this. This should not create any problem of interpretation, as long as the speed at which the piecework tasks were executed did not vary much in the short run.

2) Industrial production data are from the Federal Reserve Board. See "New Federal Reserve Index of Industrial Production," *Federal Reserve Bulletin*, August 1940, pp. 753–69 and 825–74.

3) Wholesale price indexes are from the Bureau of Labor Statistics. See the following

publications of the U.S. Department of Labor: *Handbook of Labor Statistics* (Bulletin 541, 1931; Bulletin 616, 1936; Bulletin 694, 1941), and *Wholesale Prices 1913 to 1927* (Bulletin 473, 1929). For the automobile industry I merged two BLS series of motor vehicles prices. Neither series covered 1935; the price series on all metal products was used to interpolate the automobiles price series for that year.

4) The consumer price series is from Sayre (1948).

5) The *NRA* dummy is set equal to one for all months from September 1933, when the first NRA industry codes went into effect, until May 1935, when the Act was declared unconstitutional. The monthly data on man-days idle due to strikes (used in the construction of the *UNIONPOWER* variable) are from the Bureau of Labor Statistics (Bulletins 651 and 694). The data series for total federal emergency workers, which include the WPA, the CCC, and other programs, is from the National Industrial Conference Board (NICB) *Economic Almanac* for 1941–42.

The span of the sample is January 1923 to December 1939. Although some of the data exist before 1923, there are two major problems with extending the sample further back. 1) Some of the industrial production data are missing and cannot be constructed. 2) There is a six-month gap in the NICB survey in 1922. The December 1939 stop date was chosen so as to avoid consideration of the many special features of the wartime economy.

The variables treated as endogenous and the additional instruments used in estimation in the principal equations are as follows. 1) *Demand for workers equation*: PAY_t and $QADJ_{t+1}$ are taken to be endogenous. ($QADJ_{t+1}$ is treated as endogenous because of the measurement error problem created when a future value of a variable is used in place of a forecast. See the text.) Additional instruments are $QADJ_{t-1}$, HRS_{t-1}, $UNIONPOWER_t$, and the current value and two lags of the cost-of-living variable *COL*. 2) *Demand for hours of work equation*. The endogenous variable is the cost variable, $HCOST_t$. Additional instruments are EMP_{t-1}, $UNIONPOWER_t$, and the current value and two lags of *COL*. 3) *Earnings equation*. Endogenous variables are EMP_t and HRS_t. Instruments were the current and two lagged values of production Q and current and two lagged values of *COL*. Because it was observed earlier that current employment was highly correlated with one-month-ahead production in some industries, I also used as an instrument a forecast of one-month-ahead production based on a univariate autoregression. 4) *Dynamic labor supply, or "Darby," equation*. Endogenous variable is $EARN_t$. Instruments are the same as in the earnings equation above.

REFERENCES

Altonji, Joseph, "The Intertemporal Substitution Model of Labour Market Fluctuations: An Empirical Analysis," *Review of Economic Studies*, Special Issue 1982, *49*, 783–824.

______ **and Ashenfelter, Orley,** "Wage Movements and the Labour Market Equilibrium Hypothesis," *Economica*, August 1980, *47*, 217–45.

Ashenfelter, Orley, "Unemployment as Disequilibrium in a Model of Aggregate Labor Supply," *Econometrica*, April 1980, *48*, 547–64.

Baily, Martin N., "The Labor Market in the 1930's," in James Tobin, ed., *Macroeconomics, Prices, and Quantities*. Washington: The Brookings Institution, 1983.

Bakke, E. Wight, *The Unemployed Worker: A Study of the Task of Making a Living Without a Job*, New Haven: Yale University Press, 1940.

Ball, Robert J. and St. Cyr, E. B. A., "Short-Term Employment Functions in British Manufacturing Industry," *Review of Economic Studies*, July 1966, *33*, 179–207.

Beney, M. Ada, *Wages, Hours, and Employment in the United States, 1914–1936*, New York: National Industrial Conference Board, 1936.

Bernanke, Ben, "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, June 1983, *73*, 257–76.

______, "Employment, Hours, and Earnings in the Depression: An Analysis of Eight

Manufacturing Industries," Working Paper No. 1642, NBER, June 1985.

______ and **Powell, James,** "The Cyclical Behavior of Industrial Labor Markets: A Comparison of the Pre-War and Post-War Eras," in R. J. Gordon, ed., NBER Conference Volume on Business Cycles, forthcoming.

Bernstein, Irving, *The Lean Years: A History of the American Worker, 1920–1933*, Boston: Houghton-Mifflin, 1960.

Brechling, Frank P. R., "The Relationship Between Output and Employment in British Manufacturing Industries," *Review of Economic Studies*, July 1965, *32*, 187–216.

Bry, Gerhard, "The Average Workweek as an Economic Indicator," Occasional Paper No. 69, NBER, 1959.

Creamer, Daniel, "Behavior of Wage Rates During Business Cycles," Occasional Paper No. 34, NBER, 1950.

Darby, Michael R., "Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–41," *Journal of Political Economy*, February 1976, *84*, 1–16.

Daugherty, Carroll R. et al., *The Economics of the Iron and Steel Industry*, New York: McGraw-Hill, 1937.

Davis, Horace B., *Shoes: The Workers and the Industry*, New York: International Publishers, 1940.

Fabricant, Solomon, *The Output of Manufacturing Industries, 1899–1939*, NBER, New York: Arno Press, 1940.

______, *Employment in Manufacturing, 1899–1939*, NBER, New York: Arno Press, 1942.

Fair, Ray C., *The Short-Run Demand for Workers and Hours*, Amsterdam: North-Holland, 1969.

Feldstein, Martin, "Specification of the Labour Input in the Aggregate Production Function," *Review of Economic Studies*, October 1967, *34*, 375–86.

Ireland, N. J. and Smyth, D. J., "Short-Term Employment Functions in Australian Manufacturing," *Review of Economics and Statistics*, November 1967, *49*, 537–44.

Lebergott, Stanley, *Manpower in Economic Growth: The American Record Since 1800*, New York: McGraw-Hill, 1964.

Lucas, Robert E., Jr., "Capacity, Overtime, and Empirical Production Functions," *American Economic Review Proceedings*, May 1970, *60*, 23–27.

______, "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______ and **Rapping, Leonard A.,** "Real Wages, Employment, and Inflation," *Journal of Political Economy*, September/October 1969, *77*, 721–54.

______ and ______, "Unemployment in the Great Depression: Is There a Full Explanation?," *Journal of Political Economy*, January/February 1972, *80*, 186–91.

Moore, Geoffrey H., "Business Cycles and the Labor Market," *Monthly Labor Review*, March 1955, *78*, 288–92.

Muth, John F., "Optimal Properties of Exponentially Weighted Forecasts," in Robert E. Lucas, Jr. and Thomas Sargent, *Rational Expectations and Econometric Practice*, Minneapolis: University of Minneapolis Press, 1981.

Nadiri, M. Ishaq and Rosen, Sherwin, *A Disequilibrium Model of Demand for Factors of Production*, NBER, New York: Columbia University Press, 1973.

Plessner, Yakir and Yitzhaki, Shlomo, "Unemployment and Wage Rigidity: The Demand Side," *Oxford Economic Papers*, July 1983, *35*, 202–12.

Quandt, Richard E. and Rosen, Harvey S., "Unemployment, Disequilibrium and the Short-Run Phillips Curve: An Econometric Approach," Working Paper No. 1648, NBER, June 1985.

Rees, Albert, "On Equilibrium in Labor Markets," *Journal of Political Economy*, March/April 1970, *78*, 306–10.

______, "Real Wages and Inflation: Rejoinder," *Journal of Political Economy*, January/February 1972, *80*, 192.

Rosen, Harvey S., and Quandt, Richard E., "Estimation of a Disequilibrium Aggregate Labor Market," *Review of Economics and Statistics*, August 1978, *60*, 371–79.

Rosen, Sherwin, "Short-Run Employment Variation on Class-I Railroads in the U.S., 1947–1963," *Econometrica*, July/October 1968, *36*, 511–29.

Rotemberg, Julio, "Sticky Prices in the United

States," *Journal of Political Economy*, December 1982, *90*, 1187–211.

Sargent, Thomas, "Estimation of Dynamic Labor Demand Schedules Under Rational Expectations," *Journal of Political Economy*, December 1978, *86*, 1009–44.

Sayre, R. A., "Wages, Hours, and Employment in the United States, 1934–1939," *Conference Board Economic Record*, March 28, 1940, *2*, 115–37.

______, *Consumers' Prices, 1914–1948*, New York: National Industrial Conference Board, 1948.

Stockman, Alan C., "Aggregation Bias and the Cyclical Behavior of Real Wages," research paper, University of Rochester, 1983.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

Yellen, Janet C., "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, *74*, 200–05.

Tests of the Rational Expectations Hypothesis

By MICHAEL C. LOVELL*

This paper reviews evidence from a number of empirical studies challenging the validity of the received hypothesis of rational expectations. The analysis does not rest primarily on new empirical evidence, but instead on evidence from a number of studies; some recent, some unpublished, many of older vintage. I demonstrate that the cumulative empirical evidence does not establish that the received doctrine of rational expectations dominates alternative hypotheses about expectations. After reviewing the models and the evidence, I express my qualms about the dangers stemming from the categorical acceptance of the rational expectations hypothesis as "stylized fact" to the exclusion of alternatives in empirical investigations, in theoretical research, and, most important, in policy analysis.

Is it appropriate to test the rational expectations hypothesis at the micro level? It can be said that the force behind the rational expectations argument had nothing to do with claims concerning its empirical validity. Thomas Sargent argues:

> Research in rational expectations and its dynamic macroeconomics has a momentum of its own. That momentum stems from the logical structure of rational expectations as a modeling strategy, the questions that it invites researchers to face, and the standards that it imposes for acceptable answers to those questions. [1982, p. 382]

Sargent does not claim that the assumption that expectations are rational is realistic; he does not claim that the momentum of rational expectations derives from direct empirical evidence.[1]

Edward Prescott has argued that the rational expectations hypothesis is not amenable to direct empirical test: "Like utility, expectations are not observed, and surveys cannot be used to test the rational expectations hypothesis. One can only test if some theory, whether it incorporates rational expectations or, for that matter, irrational expectations, is or is not consistent with observations" (1977, p. 30).

It seems to me that it may be a mistake to argue that we can divide variables into those that are observable and those that are not. After all, utility can be measured up to a linear transformation; measuring sales expectations, while not easy, may be no more difficult than trying to measure economic profit. A theory that claims to have a strong microeconomic foundation should be amenable to testing with micro data. Observe that over the last several decades a number of economists—from Franco Modigliani and Owen Sauerlander (1955) to Otto Eckstein, Patricia Mosser, and Michael Cebry (1984)—has found that survey observations on expectational variables can be of assistance in the empirical modeling of economic behavior and econometric forecasting.

*Department of Economics, Wesleyan University, Middletown, CT 06457. Drafts of an earlier version, "Inventories and Rational (?) Expectations" were presented at a joint International Society for Inventory Research–Canadian Economic Association Session at the annual meetings of the CEA, Guelph, May 1984, and at the Third International Symposium on Inventory Research, Budapest, August 1984. I am indebted to M. Burstein, A. L. Levine, B. J. Moore, J. Muth, and F. Rozwadowski for helpful comments on the earlier drafts.

[1]While the momentum of rational expectations may derive from the fact that John Muth's rational expectations hypothesis provides a fundamental extension to the classical economic paradigm, this explanation does not account for the fact that Muth's contribution lay dormant for a number of years, including a period in which Muth and Robert Lucas were colleagues at Carnegie-Mellon; and while Thomas Sargent was exposed to the concept of rational expectations while at Carnegie-Mellon in 1967, he did not pursue the concept at the time (see Aljo Klamer, 1983, p. 61). The momentum of rational expectations may well derive as much or more from appreciation of its forceful policy implications as from its intellectual contribution to classical theory per se.

It has long been argued that theories should be judged on their predictive ability rather than on the basis of the validity of their simplifying assumptions, which must of necessity be false. In a recent paper Robert Lucas asserts: "Any model that is well enough articulated to give clear answers to the questions we put to it will necessarily be artificial, abstract, patently 'unreal'" (1980, p. 696).

In explaining how confidence in the validity of a model's predictions can be earned, Lucas goes on to state:

> ... *individual* responses can be documented relatively cheaply, occasionally by direct experimentation, but more commonly by means of the vast number of well-documented instances of individual reactions to well-specified environmental changes made available "naturally" via censuses, panels, other surveys, and the (inappropriately maligned as "casual empiricism") method of keeping one's eyes open. [p. 696]

Lucas's statement raises the issue of whether evidence on individual behavior from surveys, censuses, and so forth is admissible only for the testing of the predictions of a theory, or whether they can and should be used to test the validity of the rational expectations hypothesis itself. My own view is that the appropriate realm for empirical research should not be demarcated in terms of the dichotomy between assumptions and predictions—I think that direct testing of the rational expectations hypothesis is an appropriate and worthwhile activity.[2] In order to be able to claim that a theory is based on firm micro foundations requires more than the derivation of propositions from the assumption that economic agents maximize, however esthetically pleasing such derivations may be; a theory that is said to be based on micro foundations should survive empirical testing at the level of the individual decision making unit. To the extent that the survey evidence supports the hypothesis of rational expectations, results derived under that assumption, policy impossibility theorems, and so forth, will be both more interesting and more demanding of serious attention.

This paper examines the evidence. After reviewing in the next section the features distinguishing Muth's theory of rational expectations from alternative models, I examine in Section III the weight of the evidence accumulated to date from a variety of empirical studies.

I. On the Structure of Expectations

The rational expectations hypothesis is only one of a variety of strategies that have been used by researchers in modeling expectations; many were developed in the 1950's. I first look at the alternatives and compare the rival models; then I shall look at the weight of the accumulated empirical evidence provided by a number of studies.

A. *Ferber's Law*

Robert Ferber (1953) concluded from interviews with a number of business enterprises that, in making forecasts, firms typically attempt to allow for seasonal movements by modifying the figure for the corresponding quarter of the previous year in the light of recently observed trend; that is to say, with quarterly observations Ferber's law states:

$$(1)\quad P_t = \rho_0 + A_{t-4}\left[\rho_1 + \rho_2 (A_{t-1} - A_{t-5})/(A_{t-5})\right].$$

Here P_t is the value predicted for quarter t on the basis of lagged actual values A_{t-i}, where time is measured in quarterly units. Observe that expectations are simply "same-as-last-year" forecasts if $\rho_0 = \rho_2 = 0$ and $\rho_1 = 1$. However, expectations are last period's

[2] This is an unresolved methodological issue on which scholars can respectfully disagree. Arnold Zellner (1985, p. 258) supports the use of micro and industry data in examining relationships suggested by macroeconomic research. While James Tobin (1980, p. 29) and Herbert Simon (1979, p. 505) support direct empirical testing of the rational expectations hypothesis, Prescott is but one of a number of distinguished economists holding the opposite viewpoint.

experience modified by a crude seasonal adjustment factor if $\rho_0 = 0$ and $\rho_1 = \rho_2 = 1$; that is,

$$(1') \qquad P = A_{-1}(A_{-4}/A_{-5}).$$

On the basis of his study of the Railroad Shippers Forecast data, Ferber concluded that expectations are regressive; that is, he found that ρ_2 was considerably less than one, implying that recent gains or losses since last year are not expected to persist.[3] In an interesting study of inventory behavior and the production decision, John Johnston (1961) used Ferber's equation in order to proxy out unobservable sales anticipations.

B. *Adaptive Expectations (Exponential Smoothing)*

The "adaptive" model of expectations formation, which stems from John Hicks' (1939) concept of the elasticity of expectations, has been analyzed by Marc Nerlove (1964) and advocated as a practical procedure by management scientists, as in Charles Holt et al. (1960). In its most elementary form, this model may be written

$$(2) \qquad P = A_{t-1} + \lambda(P_{t-1} - A_{t-1}).$$

Thus, the prediction is same-as-last-period if last period's forecast turned out to be perfectly accurate. As one extreme case, if $\lambda = 0$, the model reduces to a naive prediction of no change; alternatively, if $\lambda = 1$ we continue with the same static forecast as before without revision for current error.

C. *"Implicit" vs. "Rational" Expectations*

An alternative to these two structural approaches is to avoid explicit modeling of the process by which expectations are generated. Rather, certain reasonable stochastic properties are hypothesized.

First, it has seemed reasonable to many investigators to hypothesize that expectations will be *unbiased*; that is, the expected value of the forecast error ε is zero. More formally, I define

$$(3) \qquad \varepsilon = P - A$$

and impose the restriction that $E(\varepsilon) = 0$.

Edwin Mills (1957) imposed an additional restriction in developing his concept of "*implicit expectations*" in connection with his fundamental empirical study of inventory behavior. Specifically, Mills conjectured that the prediction error is uncorrelated with the actual realization; with this restriction, the basic assumption of the regression model is satisfied with the anticipated variable selected as the dependent variable:

$$(4) \qquad P = \alpha_0 + \alpha_1 A + \varepsilon$$

$$\text{with} \quad \alpha_0 = 0; \quad \alpha_1 = 1; \quad E(\varepsilon) = 0.$$

On the basis of this argument, Mills used the actual realization as a proxy for the unobserved anticipated level of sales in his empirical study of inventory behavior.

John Muth (1961) pioneered a procedure that is just the opposite of Mills' implicit expectations hypothesis. For rational expectations, Muth required that the forecast error be distributed independently of the anticipated value; that is,

$$(5a) \qquad A = \beta_0 + \beta_1 P + \varepsilon$$

$$\text{with} \quad \beta_0 = 0; \quad \beta_1 = 1; \quad E(\varepsilon) = 0.$$

For Muth, ε must be uncorrelated with P, the predictions; therefore, it must be correlated with A, the actual realizations; hence the variance of A is larger than the variance of P. All this is precisely the reverse of Mills' implicit expectations, which have a larger variance than the actual realizations.[4]

[3] The same data source was also used in the pioneering Modigliani and Sauerlander study of inventory behavior. The validity of the Railroad Shipper Forecast data was questioned by Albert Hart (1960) and by me (1964: appendix).

[4] The sample variance of actual realizations will exceed that of the anticipations, as required by the rational expectations hypothesis, only if $r < b_1$, where b_1 is the least squares estimate of β_1 in equation (5a). To verify this, observe that the regression coefficient $a_1 = \text{cov}(a, p)/S_a^2$ while $b_1 = \text{cov}(a, p)/S_p^2$; further $r^2 = a_1 b_1$; therefore, $S_p/S_a = r/b_1$.

To earn the *fully rational* accolade more is required; specifically:

The prediction error must be uncorrelated with the entire set of information that is available to the forecaster at the time the prediction is made.

This rationality concept might be called "sufficient expectations," for it is closely related to the statistical concept of a "sufficient estimator," which may be loosely defined as an estimator that utilizes all the information available in the sample.

One implication of this requirement is that the prediction error must be uncorrelated with historical information on prior realizations of the variable being forecast; this *weak rationality* condition, as it is sometimes called, implies that if lagged values of A are added to the right-hand side of regression model (5a), they must appear with zero coefficients; for example, the coefficient b_2 in the following regression should not differ significantly from zero:

$$A = b_0 + b_1 P + b_2 A_{t-1}. \tag{5b}$$

The "full rationality" conjecture also has a more demanding implication: it requires as a condition of "*strong rationality*" that any other variables known to the forecaster (for example, public information on the rate of growth of the money supply, federal deficits, and the unemployment rate) must also be uncorrelated with the forecast error.

D. *Change Underestimation*

In empirical work on the determinants of inventory investment (1961), I proxied out the unobserved expectational variables by invoking the conjecture that the predicted change is a fraction ρ of observed changes:

$$P - A_{t-1} = \rho(A - A_{t-1}) + \varepsilon. \tag{6}$$

This equation is less restrictive than either implicit or rational expectations. With $\rho = 1$, this model reduces to either implicit or rational expectations, depending on whether it is conjectured that ε is distributed independently of P or A. In contrast, if $\rho = 0$ and $\varepsilon = 0$, we have naive "no change" extrapolative forecasts; with $0 < \rho < 1$, there is a systematic tendency to underestimate change, as hypothesized by J. M. Keynes: "...it is sensible for producers to base their expectations on the assumption that the most recently realized results will continue except in so far as there are definite reasons for expecting a change" (1936, p. 51).

E. *Evaluation*

In my judgement the choice among these alternatives is not easily resolved by a simple appeal to maximization or theoretical principle. Although closely related concepts, the choice between the mutually exclusive rational and implicit expectations models is not easy. One or the other of two opposing arguments can be advanced to rationalize the alternative formulations:

1) First, suppose the sales forecaster makes the prediction on the basis of a regression model (for example, historical sales explained by earlier values of sales and other variables); then the prediction error (at least over the sample period) will be uncorrelated with the explanatory variables. For example, suppose the forecaster fits to historical data the regression equation

$$A = b_0 + b_1 A_{-1} + b_2 A_{-4} + e \tag{7}$$

and uses the resulting least squares coefficients to generate predictions

$$P = b_0 + b_1 A_{-1} + b_2 A_{-4}, \tag{8}$$

the resulting prediction errors will be uncorrelated with the explanatory variables, at least over the sample period, in accordance with Muth's rationality hypothesis.

2) Alternatively, with regard to Mills' concept of implicit expectations, it was pointed out by Albert Hirsch and me (1969, pp. 73–74) that sales forecasts derived by periodically surveying a sample of reliable customers are likely to satisfy this condition; that is, a survey of a random sample of customers will yield an estimate of average sales per customer which will be subject to sampling error; the survey results will be randomly distributed about the actual popu-

lation response, as required by Mills' implicit expectations model.[5]

For certain econometric purposes it facilitates matters to assume implicit rather than rational expectations. Consider the problem of estimating the standard flexible accelerator inventory model I used earlier (1961):

$$(9)\quad I_t = \delta\gamma_0 + (1+\delta\gamma_1)X_t^e + (1-\delta)I_{t-1} - X_t.$$

Here I_t is end of period inventories, X_t^e is anticipated sales, and X_t actual sales.[6] Substituting from equation (4) yields

$$(10)\quad I_t = \delta\gamma_0 + \delta\gamma_1 X_t + (1-\delta)I_{t-1} + (1+\delta\gamma_1)\varepsilon_t.$$

As Mills explained (1957), because the implicit expectations forecast error ε_t is uncorrelated with X_t, the application of least squares to the equation obtained by using the implicit expectations proxy will yield asymptotically unbiased parameter estimates. With rational expectations, the estimation problem is more complex. Thus an advocate of the principle of parsimony might cite Occam's Razor in support of implicit over rational expectations.

In support of the rational expectations hypothesis, it should be observed that it is precisely this concept of expectations that is required in order for the "certainty equivalence" argument of Herbert Simon (1956) and Henri Theil (1957) to go through in the derivation of optimal linear decision rules. A classic management science application has to do with the task of optimally scheduling production, as formalized in Holt et al. A firm interested in utilizing their linear decision rule procedure for production scheduling *should* use forecasts that satisfy the rational expectations condition that the forecast error be uncorrelated with the forecast. While it might be tempting to conjecture that this procedure may have been put into practice by the generations of graduate business school students nurtured on the production smoothing algorithm at Carnegie-Mellon's Graduate School of Industrial Administration and elsewhere, caution is required. The Holt et al. argument does not suffice to establish the adequacy of the rational expectations formulation, for quite restrictive assumptions are required in order for the certainty equivalence argument to go through. The loss function must be quadratic, and there must be no sign constraints on the decision variables (for example, negative output and inventories must be admissible); as Simon (1979) cautioned in his Nobel Laureate address, single-valued forecasts applied with linear decision rules are unlikely to suffice in more complex situations.

The notion of certainty equivalence does not necessarily go through even when the loss function is quadratic. As one example, in the Tobin-Markowitz portfolio selection model more than point estimates are required; that is, the notion of certainty equivalence does not go through because the variance of the loss depends on the decision; this contrasts with the Holt et al. production scheduling model, for they assumed that the costs of uncertainty were independent of the production decision.

As a second example, consider Milton Friedman's formal demonstration (1953) of the intuitively reasonable proposition that efforts at applying macroeconomic stabilization policy should be less aggressive when the policymaker's forecasts are subject to greater error. It turns out that Friedman's

[5]This is not the only rationalization of Mills' approach. In their empirical study of the informational content of prices in dealer securities markets, it is assumed by K. D. Garbade et al. (1979, p. 52) that a dealer's observed offering price is randomly distributed about the true equilibrium price, rather than the reverse. Although they do not mention the concept, they are assuming that the offering price is an implicit rather than rational forecast of the equilibrium outcome; they do point out, however, that their assumption will not hold when a dealer wishes to adjust his inventory position in a security.

[6]The flexible accelerator inventory model, frequently employed in empirical research, assumes that sales are exogenous. This assumption underlies much of the management science prescriptive literature on inventories and the production decision. The same assumption is employed in the Jorgenson neoclassical model of fixed investment. In contrast, Michael Brennan's (1959) approach of having the firm determine its inventory stocks on the basis of seasonal price movements is particularly relevant in the study of inventory holdings of agricultural commodities.

argument holds for Mills' implicit expectations, but not for Muth's rational expectations.[7]

That certainty equivalence does not go through in quite simple circumstances when the loss function is asymmetric is illustrated by a simple every day example:

It is best to get to the bus stop a few minutes in advance of the expected 8:00 *arrival time for your bus because your loss from missing the bus by one minute is greater than the cost of waiting an extra minute — that is, your loss function is not symmetric. Some commuters plan as though the expected arrival time were* 7:55; *other commuters set their clocks five minutes ahead.*

Finally, it must be observed that although the concept of implicit and rational expectations are mutually exclusive, it would be a relatively simple matter to modify expectations satisfying Mills' implicit expectations hypothesis in order to obtain transformed expectations that are rational in the sense of Muth; it is only necessary to apply a linear transformation to the implicit expectations, as with equation (9), utilizing coefficients obtained by regressing historical actual realization on the implicit expectations. However, I know of no evidence that firms customarily modify sample survey forecasts in this way.

II. Evidence

Direct evidence on these issues is provided by a variety of empirical studies, some recent but others of long standing, on the structure of expectations.[8]

A. *Manufacturers' Sales and Inventory Anticipations*

A rich body of *ex ante* evidence is provided by the quarterly *Manufacturers' Inventory and Sales Expectations* (*MISE*) *Survey* conducted by the U.S. Department of Commerce from late 1959 through 1976. Firms were asked for both short (2 month) and long (5 month) sales and inventory forecasts; they were also asked whether they regarded their inventory stocks as high, low, or about right relative to current sales. The survey evidence has been extensively analyzed by me (1967), by Hirsch and me, and by F. Owen Irvine, Jr. (1983). Hirsch and I had access to the responses of 83 firms in five industries through the fourth quarter of 1964 as well as the industry aggregates.

Hirsch and I (p. 71) reported that the sales expectations of individual firms are biased, as defined by equation (3) above, which contradicts both Muth's rational and Mills' implicit expectations models: some firms are *perennial optimists*, generally overestimating future sales, while others are *perennial pessimists*, usually understating sales volume. For 30 percent of the sampled firms, the mean of anticipated sales, two-months horizon, differed from the mean of actual realizations at the 5 percent level of significance. However, the overestimates of the optimistic firms roughly cancelled the underestimates of pessimistic firms so that for industry aggregates there is no bias; this offsetting of systematic error partially explains why the aggregates of anticipations data appear to be more accurate than the predictions of individual firms.

The rational expectations hypothesis asserts that the variance of actual realizations will exceed the variance of forecasts; the implicit expectations hypothesis holds the opposite. In fact, Hirsch and I (p. 74) found a mixed picture, for a sizable proportion of firms (about 35 percent) sales anticipations have a larger variance than realizations.

As a further test of the rational expectations model, Hirsch and I considered regressions of the form:[9]

$$(11)\quad A_t = \beta_0 + \beta_1 P_t + \beta_2 A_{t-1} + \beta_3 A_{t-4} + \varepsilon.$$

[7]And the policy formulation task is still more involved when the parameters of the structure by which policy has its impact have to be estimated (compare William Brainard, 1967; Prescott, 1971).

[8]This review will not cover survey evidence of professional forecasters, such as that provided by the Livingston Survey.

[9]As before, A_t is the actual realization and P_t is the prediction of the outcome; A_{t-1} is last quarter's realization and A_{t-4} is the same-period-last-year realization,

Under Muth's assumption that firms exploit efficiently all available information in making their forecasts, we should find $\beta_1 = 1$ and $\beta_2 = \beta_3 = 0$. On the other hand, if firms fail to fully exploit the information on last period sales or same-period-last-year sales, these conditions will be violated.

The results Hirsch and I reported (pp. 171–77) are supportive of the rational expectations hypothesis for the durable manufacturing aggregate and for the seven component industry aggregates; for nondurables, however, β_2 is substantially less than unity for a number of industries. Further, too many of the estimated values of β_3 and β_4 have large t values. And the evidence for individual firms is even more discouraging for the rational expectations hypothesis. A pooled regression was run for each of the five industries for which individual firm observations were available; almost always, the value of β_2 was significantly less than unity; equally discouraging, the remaining two regression coefficients were usually significantly different from zero, which contradicts the rational expectations hypothesis.

Why do firms fail to exploit fully the information in their own sales history in formulating their forecasts of future sales volume? One possibility, pointed out by Hirsch and me (p. 177), is that while it may be true that a decision maker who knew the parameters of equation (11) could improve the accuracy (as measured by the root mean square error) of the raw forecasts with an appropriate linear transformation, departures from the orthogonality conditions imposed by Muth may arise because the decision maker has not accumulated enough historical evidence to obtain precise estimates of the parameters of equation (11).

Hirsch and I (pp. 181–85) concluded that in empirical work the most appropriate assumption to make about expectations when anticipations are not directly observable depends on the level of aggregation. At the firm level Ferber's law and the exponential smoothing model both yield a better estimate of anticipated sales than is provided by the actual realization proxy. For industry aggregates, however, it is better to use actual sales as a proxy for anticipations rather than to assume that expectations are generated either by Ferber's law or by exponential smoothing. This discrepancy arises because the cancelling of offsetting forecasting errors of individual firms makes aggregate anticipations much more accurate predictors of aggregate realizations, and conversely. But Hirsch and I also found that the assumption that predicted changes are proportional to actual changes, the relaxation of the assumption of rational expectations restrictions on equation (11), does better than either Ferber's law or exponential smoothing is predicting short sales anticipations.

B. *Further Tests Based on the MISE Survey*

In his more recent study, Irvine utilizes the data for the entire seventeen years over which the *Survey* was conducted, 1959 through 1976, but only for the durable, nondurable, and total manufacturing aggregates rather than disaggregated to the industry or the firm level; as a result, he could not investigate the tendency observed by Hirsch and me for some firms to be perennial optimists and others perennial pessimists. Focusing on the short sales forecast aggregates, Irvine finds that there was a two-and-one-half year sequence of sales underprediction beginning in late 1972, which might arise because firms did not adequately allow for the dramatic upward sweep of inflation in pricing projected sales volume—this could be interpreted as a protracted transitional period in which business firms were slowly learning about a change in structure. For the pre-OPEC period, 1961–72, his tests revealed no evidence inconsistent with the hypothesis that the one-period-ahead sales forecasts of durable manufacturing are fully rational. But he found for the nondurable manufacturing aggregates, as had Hirsch and I, that expectations are not fully rational in that they do not appropriately incorporate information on seasonality and the rate of growth of the money supply.

which may appear significant if the forecaster fails to exploit systematic seasonal movements. The data were not seasonally adjusted.

C. *Five Pittsburgh Firms*

Muth (1985) tested alternative theories of expectations on monthly data spanning the period 1957–70 for five Pittsburgh-based business firms, two of which were steel producing, two metal fabricating, and one an electric utility. For each individual firm Muth had observations on anticipated production for three successive months plus information on realized production, deliveries, new orders, the order backlog, and total inventories. The data suggest that some firms are perennial optimists while other firms are perennial pessimists, as was the case with the firms studied by Hirsch and me.[10] In the majority of cases, the variance of anticipations is larger than the variance of the realizations, which is inconsistent with the rational expectations hypothesis.[11] Muth also found his data inconsistent with a number of alternative structural expectations models, with the possible exception of the expectations revision model of Hicks and of David Meisselman (1962). As explained later in this paper, Muth was led by the negative empirical evidence to substantially modify his original model of rational expectations.

D. *Price Expectations*

Inflationary expectations are of obvious interest in their own right; they are also of special interest in the study of inventory behavior for two reasons. First, in attempting to measure the real rate of interest, a component of inventory carrying cost, it is necessary to take into account the price changes expected by business firms; second, in times of unanticipated inflation it is particularly useful to be able to decompose errors in anticipating sales volume into errors in predicting real sales volume and errors in estimating sales price.[12]

Two studies by Frank de Leeuw and Michael McKelvey (1981, 1984) exploit the evidence on the price expectations of business firms provided by the responses to the year-end survey of Business Expenditures on Plant and Equipment conducted by the Bureau of Economic Analysis since 1970. They report (1981, p. 302) that the expected price changes are somewhat more accurate than simply forecasting the same rate of inflation as last year; specifically, Theil's U-statistic, the ratio of the root mean square error of the observed forecast over the root mean square error that would be made by a forecaster who always predicted the same inflation rate as last year, averages out to about 77 percent over the decade of the 1970's, ranging from a most impressive 58 percent for textiles to a tie at 100 percent for food and beverages. They found that the two rounds of OPEC price hikes caused major errors in the anticipated prices of goods and services sold, the first in 1974 and the second in 1981. The expected percent change in price in 1974 was only 5.3 percent while the actual hike was 16 percent; but for 1975 the expected 8.8 percent inflation fell just short of the actual 8.9 percent. This suggests that at least part of the two-and-one-half-year sequence of underprediction of nominal sales reported by Irvine can be attributed to unanticipated inflation.

De Leeuw and McKelvey report on a number of pooled industry cross-section time-series regressions testing the rational expectations hypothesis. When they regressed the actual rate of increase in the sales price (p_t) on the anticipated change (p_t^e) over the period 1971–80, the regression suggested by equation (5a) above, they obtained

$$(12) \qquad \underset{(1.12)}{p_t = -0.112} + \underset{(.155)}{1.345 p_t^e} \qquad \bar{R}^2 = .304.$$

An F-test of the joint hypothesis $\beta_0 = 0.0$ and $\beta_1 = 1.0$ yields a highly significant F-

[10] Of 24 bias t-ratios, 12 were greater than 2 in magnitude; all but 4 were greater than unity.

[11] In half of 28 cases, the variance of the forecast was larger than the variance of the realization.

[12] The evidence considered here, in contrast to the Livingston Survey data analyzed by such writers as John Carlson (1975), concerns the expectations of business firms and consumers rather than professional forecasters.

statistic of 12.2.[13] With capital goods prices, the results were also inconsistent with the rational expectations hypothesis, but with slope coefficients significantly *less* than unity. Equally serious, de Leeuw and McKelvey present evidence suggesting that firms do not fully exploit publicly available information on the lagged rate of growth of the money supply and capacity utilization.

In the follow-up study (1984), de Leeuw and McKelvey worked with both individual firm data and with data grouped in order to mitigate the problem of errors in the variables. They generally found that regressions of the form (5a) violate the rationality hypothesis, the estimates of β_1 being substantially less than unity on both grouped and firm disaggregated data. Expected price change is determined by a variety of variables, including lagged expected rates of inflation and recently observed changes in the rate of price change. They are able to conclude, however, that expectations may not be subject to long-run bias (i.e, $E(\varepsilon)=0$ in equation (3) in the long run).

A study by Edward Gramlich (1983) based on a quite distinct body of price expectations data provided similar results. Using time-series derived from the University of Michigan household survey data, he found a slope coefficient of 1.222, only slightly flatter than the sales price slopes reported by de Leeuw and McKelvey (equation (12) above).

E. *Wage Expectations*

Jonathan Leonard (1982) has analyzed data on employers' wage expectations provided by the Endicott survey on average starting wages for inexperienced college graduates. Data for eight occupational categories are collected from a sample of 170 large and medium-sized corporations. Expectations appear to be biased downward, for in each of the eight occupational categories employers underestimate the wages they will have to pay new recruits. The slope coefficients in equation (5a) are closer to the value of unity than those obtained in most other studies; nevertheless, he finds in contradiction to the hypothesis of rational expectations that the F-test is significant in five of eight occupational categories. He concludes that firms underestimate wages because they underestimate demand; forecast errors are not explained by misperceptions of inflation or by either the expected or the unanticipated money supply change.

F. *Data Revisions*

Business analysts and economic forecasters are confronted with a confusing sequence of flash, preliminary, provisional and revised numbers for each observation of interest. It has been observed by a number of writers, most notably Arnold Zellner (1958) and Rosanne Cole (1970), that preliminary data on *GNP* and other economic indicators deviate systematically from subsequent revisions. A review of the evidence suggests that the official preliminary data are *not* always rational predictors of the revised time-series that eventually appear.[14] Such departures are of importance if economic agents utilizing official information sources as part of the information set in making their own projections are induced to make forecasts that fail to satisfy the rational expectation hypothesis.

Evidence that preliminary X-11 seasonally adjusted money supply growth is subject to systematic error has appeared in a number of studies.[15] As a result, a revised X-11 ARIMA seasonal adjustment procedure was adopted in 1982. That the revised seasonal adjustment procedure yields preliminary data that

[13] Leaving out the 1974 OPEC shock year yields results that are even more damaging to the rational expectations hypothesis.

[14] While revisions in official data series do occasionally earn comment, the possibility of improving preliminary figures through appropriate linear transformation when they are not rational forecasts of the revision has apparently escaped attention from professional economists and economic analysts. However, it is reported that Edward J. Hyman, chief economist at the New York securities firm of Cyrus J. Lawrence Inc., attempts to forecast the business cycle by tabulating the direction of revision of a number of economic indicators (*Wall Street Journal*, June 6, 1984, p. 35).

[15] See the literature cited by Scott Hein and Mack Ott (1983, pp. 19).

do not constitute rational predictions of the revised data is suggested by a regression presented by Scott Hein and Mack Ott:

$$(13)\quad M1G^{revised} = \underset{(1.480)}{3.149} + \underset{(0.126)}{0.581}\, M1G^{Preliminary}$$

$$R^2 = 0.681;\ D\text{-}W = 2.26.$$

This is precisely the form of the rational expectations equation (5a) above, with the preliminary data on the rate of growth of the money supply ($M1G^{Preliminary}$) serving as the predictor of the revised figure ($M1G^{revised}$). As Hein and Ott point out, the coefficient on the preliminary growth rate of the money supply deviates significantly at the 5 percent level from unity; equally serious, the intercept deviates significantly from 0.

Hein and Ott do not mention certain interesting implications of their results. First of all, rejecting either one of these two null hypotheses would suffice to establish that the preliminary data are *not* rational forecasts of the revised seasonally adjusted data. In contrast, a similar regression shows that preliminary nonseasonally adjusted data provide a rational forecast of the revised seasonally unadjusted rate of money supply growth. It is also of interest to observe that if the regression is run in the opposite direction, in accordance with Mills' implicit expectations assumption of equation (4) above, the regression coefficient is 1.17; since this is rather close to unity, the new X-11 ARIMA procedure appears to be generating *implicit* rather than *rational* preliminary forecasts of the revised rate of money supply growth.[16]

Evidently, the problem of achieving rational forecasts by learning from prior experience has not been solved by the many statisticians who have been working on the problem of seasonal adjustment—and if the statisticians are slow learners, it is hard to believe that individual business enterprises devoting fewer resources to the problem are likely to learn from their own historical experience how to achieve rational sales forecasts within a reasonable time frame. To the extent that preliminary seasonally adjusted money supply growth rates are taken seriously by economic agents, departures from rationality may result from the failure of the estimates to satisfy Muth's concept of rational expectations. The greater variance of the preliminary data, the numbers that everyone looks at, may help to explain the excessive variability in financial markets that has been observed by Robert Shiller (1978).

G. *Government Forecasts*

Consider two forecasts provided by the federal government: budget revenue, and EPA mileage estimates. Are these forecasts rational? If not, consumers of such forecasts may make decisions that are inconsistent with the rationality postulate.

Each January the Treasury Department estimates tax receipts for the coming fiscal year. An examination of data assembled for the period 1963–78 by the Congressional Budget Office (1981) reveals that while the estimates are often imprecise, they are not significantly biased.[17] Specifically, actual revenue averaged $218.67 billion over this period, just slightly in excess of the average estimate of $217.32 billion; while some of the forecast errors are of substantial magnitude, the standard deviation of the forecast error being $9.22 billion, the slight underestimate of $1.35 is not significant; there is no reason to conclude that the tax revenue estimates are subject to systematic bias. Further, regressing the actual realization on the forecast yields Actual Revenue (AR)

$$(14)\quad AR = \underset{(6.315)}{-1.041} + \underset{(0.028)}{1.009}\, Forecast + e$$

$$\bar{R}^2 = 0.99;\quad D\text{-}W = 2.004.$$

[16]In contrast, similar regressions reported by Hein and Ott for the rate of growth of the *nonseasonally adjusted* money supply appear to be compatible with both the rational and the implicit expectations hypothesis. Seasonally adjusted unemployment rates may not be rational forecasts of the revised series.

[17]Although the Congressional Budget Office examined the accuracy of the forecasts, they did not test for rationality.

The slope coefficient differs insignificantly from unity; the Treasury forecasts satisfy the orthogonality requirement imposed by Muth's theory of rational expectations.[18]

The Environmental Protection Agency publishes auto mileage estimates that are designed to help car buyers compare the relative fuel efficiency of different models. Do these predictions, based on stationary 23-minute exhaust emission dynamometer simulations conducted by the manufactures on prototype models, constitute rational forecasts of the mileage that purchasers will realize under actual driving conditions? My comparison of the EPA estimates with Consumer Union (CU) on the road experience involving a mix of city driving, expressway driving, and driving on a 195-mile test trip revealed that contrary to conventional wisdom, the published EPA estimates for 1984 were not subject to significant optimistic bias.[19] However, the (between model) variance of the EPA forecasts substantially exceeded the mileage experienced by CU test drives, implying implicit rather than rational expectations. Further, the regression of the realization on the forecast had a slope deviating significantly from the value of unity implied by the rational expectations hypothesis:[20]

$$\text{(15)} \quad AveGal = \underset{(1.753)}{7.952} + \underset{(0.061)}{0.693}\, EPA + e; \qquad \bar{R}^2 = 0.74.$$

Evidence that the EPA forecasts do not fully incorporate the information set is provided by the following regression:[21]

$$\text{(16)} \quad AveGal = \underset{(5.349)}{22.008} - \underset{(0.001)}{0.002}\, Weight - \underset{(0.708)}{2.760}\, stan/auto + \underset{(1.413)}{3.280}\, G/D + \underset{(0.097)}{0.415}\, EPA + e \qquad \bar{R}^2 = 0.82.$$

Thus the EPA mileage estimates fail the rational expectations test in that they do not fully incorporate all the information available at the time the prediction is made. To the extent that consumers rely on the EPA fuel estimates, their purchase decisions will not be guided by rational expectations. While the last regression reveals that the EPA forecasts are not rational; it also shows that these forecasts do contain information that would contribute to improved prediction if it were used in conjunction with the other variables in the regression.[22]

IV. Implications—Should the Facts Be Allowed to Spoil a Good Story?

My survey of a number of empirical studies of expectations is not supportive of the commonly invoked rational expectations hypothesis. Quite the contrary, if the cumulative evidence is to be believed, we are compelled to conclude that expectations are a rich and varied phenomenon that is not adequately captured by the concept of rational expectations; while the predictions of some forecasters may be characterized as rational, in other instances the assumption of rationality is clearly violated. Nevertheless, there are, I think, two important reasons that can be advanced for suspending judgement on the

[18]Running this regression in the opposite direction reveals that the data are also compatible with Mills' concept of implicit expectations. When the regression was run in terms of percentage change from the preceding year, the regression results were consistent with the rational expectations concept.

[19]See my paper (1984). The shift in 1985 to the mandatory reporting of city and expressway estimates is likely to have introduced an optimistic bias that had been absent when auto stickers presented a single overall mileage figure that was based entirely on the city driving simulation.

[20]*AveGal* is the actual experience on CU road tests; *EPA* is the published EPA estimate.

[21]*Weight* is the vehicle gross weight in pounds, *stan/auto* is a dummy variable coded zero for standard and 1.0 for automatic transmission, and *G/D* is a dummy coded zero for gas and 1.0 for diesel power.

[22]Allan Murphy and Robert Winkler (1984) imply that the U.S. Weather Bureau's Model Output Statistic procedure improves predictive accuracy by using a regression-based correction equation to improve the raw model-based forecasts.

validity of tests of the rational expectations hypothesis:

First, there is the problem of measurement error. While direct empirical studies of the validity of the rational expectations hypothesis have not generated much in the way of response from rational expectations theorists, Finn Kydland and Edward Prescott did comment as follows on the results Hirsch and I reported:

> ...there may be biases in their measurement of expectations, and these biases are related to lagged sales. This is not implausible, given the subtleness of the expectations concept and the imprecision of survey instruments. Further, even if there were a systematic forecast error in the *past*, now that the Hirsch and Lovell results are part of agents' information sets, future forecast errors should not be subject to such biases. [1977, p. 479]

Is it not conceivable that further research may reveal that much of the apparent discrepancy between the rational expectations model and the evidence currently available does indeed arise from measurement error? As is well known, the presence of measurement error in the explanatory variable means that

$$\text{plim}(b) = (1+\eta)^{-1}\beta, \tag{17}$$

where η is the ratio of the variance of the measurement error to the variance of ε, provided that the measurement error is distributed independently of the true values of the explanatory variables and ε (see Johnston, 1984). Thus one might cite the downward bias generated by errors in observing expectations to explain why Hirsch and I often obtained slope coefficients significantly less than unity in regressing actual on anticipated sales. But the errors of measurement argument cuts both ways. If the errors of expectations argument is to be invoked to explain why the slope is too small in these studies, then it must also make it all the more difficult to explain the too high slope estimates obtained for the price expectations data by de Leeuw and McKelvey (1981, 1984) and by Gramlich.

Second, it must also be observed that departures from rationality may be a transient phenomenon arising because economic actors are learning to adapt to a shift in regimes; in an evolving environment more complicated tests may be required in order to determine whether satisficing or optimal learning is taking place.

A. *Muth's Errors in the Variables Reformulation*

While there may be a variety of arguments for resisting the implications of the empirical evidence, Muth was led by the evidence provided by his own and other empirical studies to fundamentally modify his original model (1985). In his new "errors in the variables" model, Muth relaxed a key restriction of his original rational expectations hypothesis. He specified

$$A_t = \alpha_t + \varepsilon_t; \quad P_t = \alpha_t + \xi_t. \tag{18}$$

Here α_t is the unobservable deterministic factor and ε_t and ξ_t are unobserved stochastic distrurbances subject to the restriction

$$E(\varepsilon_t) = E(\xi_t) = E(\alpha_t\varepsilon_t) = E(\alpha_t\xi_t) = 0. \tag{19}$$

Muth's new formulation reduces to his rational expectations model when the restriction $\sigma_\xi = 0$ holds; it reduces to Mills' implicit expectations model when $\sigma_\varepsilon = 0$.

The new model, it seems to me, can be interpreted in the following way. The stochastic term ξ_t arises from a less than full understanding of underlying deterministic forces, α_t. As in the implicit forecast model of Mills, this term could result from sampling error, as when a manufacturer relies on a sales forecast obtained from a market research survey. The other stochastic term, as in Muth's earlier rational expectations model, reflects random developments between the time the forecast is made and the actual realization.

As Muth points out, his generalization allows the variance of the predictions to exceed the variance of the actual realizations.

This means that his new model allows the slope coefficient when realizations are regressed on anticipations to be substantially less than unity, not because of errors in measuring expectations but because of an additional random element in the process by which actual anticipations are generated.

V. Conclusions

In conclusion, it seems to me that the weight of empirical evidence is sufficiently strong to compel us to suspend belief in the hypothesis of rational expectations, pending the accumulation of additional empirical evidence. This means that at this juncture three research strategies deserve more attention than they currently receive.

1) First, it is a mistake in empirical research on such phenomena as inventories to proceed under the maintained hypothesis that expectations are rational. Instead, it should be recognized that there are several competing hypotheses. Because no single hypothesis is preeminent, the researcher must test the sensitivity of empirical results on such issues as whether interest rates influence inventory holdings by reporting what happens when alternative assumptions about the structure of expectations are considered.

2) Second, more attention needs to be given to the empirical testing of the rational expectations hypothesis against its alternatives. Unfortunately, *ex ante* evidence is sparse; more resources should be devoted to the collection and dissemination of survey results. It is particularly unfortunate that the *Inventory and Sales Expectations Survey* pioneered by Murray Foss at the Department of Commerce has been discontinued.

3) Third, it would be constructive in developing theoretical models to determine how robust policy conclusions are to departures from expectational rationality. To illustrate, one can ask whether the policy conclusions derived from a model require that individual decision makers formulate their expectations rationally, or only that the aggregate of expectations held by individuals satisfy certain rationality conditions. One can also ask whether the conclusions derived under the assumption of rational expectations also go through with alternative assumptions, such as Mills' implicit expectations, Ferber's law, or a systematic tendency to underestimate change.

REFERENCES

Brainard, William, "Uncertainty and the Effectiveness of Policy," *American Economic Review Proceedings*, May 1967, *57*, 411–25.

Brennan, Michael J., "A Model of Seasonal Inventories," *Econometrica*, April 1959, *27*, 228–44.

Carlson, John, "Are Price Expectations Normally Distributed," *Journal of the American Statistical Association*, December 1975, *70*, 749–54.

Cole, Rosanne, *Errors in the Provisional Estimates of Gross National Product*, NBER *Studies of Business Cycles*, No. 21, University Microfilms, 1970.

de Leeuw, Frank, "Inventory Investment and Economic Instability," *Survey of Current Business*, December 1982, *62*, 23–31.

______ **and McKelvey, Michael J.,** "Price Expectations of Business Firms," *Brookings Papers on Economic Activity*, 1:1981, 299–314.

______ **and** ______, "Price Expectations of Business Firms: Bias in the Short and Long Run," *American Economic Review*, March 1984, *74*, 99–110.

Eckstein, Otto, Mosser, Patricia and Cebry, Michael, "The DRI Market Expectations Model," *Review of Economics and Statistics*, May 1984, *66*, 181–91.

Ferber, Robert, *The Railroad Shippers Forecasts*, Urbana: Bureau of Economic and Business Research, University of Illinois, 1953.

Friedman, Milton, "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Garbade, K. D., Pomrenze, J. L. and Silber, W. L., "On the Informational Content of Prices," *American Economic Review*, March 1979, *69*, 50–59.

Gramlich, Edward M., "Models of Inflation Expectations Formation: A Comparison

of Household and Economic Forecasts," *Journal of Money, Credit and Banking*, May 1983, *15*, 155–73.

Hart, Albert G., "Quantitative Evidence for the Interwar Period on the Course of Business Expectations: A Reevaluation of the Railroad Shippers' Forecasts," in *The Quality and Economic Significance of Anticipations Data*, Universities-National Bureau Conference, No. 10, University Microfilms, 1960.

Hein, Scott E. and Ott, Mack, "Seasonally Adjusting Money: Procedures, Problems, Proposals," *Federal Reserve Bank of St. Louis Review*, November 1983, *65*, 16–25.

Hicks, J. R., *Value and Capital*, London: Oxford University Press, 1939.

Hirsch, Albert and Lovell, Michael, *Sales Anticipations and Inventory Behavior*, New York: Wiley & Sons, 1969.

Holt, Charles C. et al., *Planning Production, Inventories and Work Force*, Englewood Cliffs: Prentice Hall, 1960.

Irvine, F. Owen, Jr., "Direct Tests of the Hypothesis that Expectations are Rational," manuscript, May 1983.

Johnston, John, *Econometric Methods*, New York: McGraw-Hill, 1984.

______, "An Econometric Study of the Production Decision," *Quarterly Journal of Economics*, May 1961, *75*, 234–61.

Keynes, J. M., *General Theory of Employment, Interest, and Money*, New York: Harcourt Brace, 1936.

Klamer, Aljo, *Conversations with Economists: New Classical Economists and Opponents Speak Out on the Current Controversy in Macroeconomics*, Totowa: Littlefield, Adams, Rowman & Alanheld, 1983.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–92.

Leonard, Jonathan S., "Wage Expectations in the Labor Market: Survey Evidence on Rationality," *Review of Economics and Statistics*, February 1982, *64*, 157–61.

Lovell, Michael C., "Manufacturers' Inventories, Sales Expectations, and the Acceleration Principle," *Econometrica*, July 1961, *29*, 293–314.

______, "Determinants of Inventory Investment," in *Models of Income Determination*, NBER *Studies in Income and Wealth*, No. 28, University Microfilms, 1964.

______, "Sales Anticipations, Planned Inventory Investment, and Realizations," in Robert Ferber, ed., *Determinants of Investment Behavior*, New York: Columbia University Press, 1967, 537–80.

______, "EPA & CU MPG Estimates," working manuscript, 1984.

Lucas, Robert E. Jr., "Methods and Problems in Business Cycle Theory," *Journal of Money, Credit and Banking*, November 1980, *12*, 696–715.

Meisselman, David, *The Term Structure of Interst Rates*, Englewood Cliffs: Prentice Hall, 1962.

Mills, Edwin S., "The Theory of Inventory Decisions," *Econometrica*, April 1957, *25*, 222–38.

______, *Prices, Output and Inventory Policy*, New York: Wiley & Sons, 1962.

Modigliani, Franco and Sauerlander, Owen H., "Economic Expectations and Plans in Relation to Short-Term Economic Forecasting," in *Short-Term Economic Forecasting*, NBER *Studies in Income and Wealth*, No. 17, New York: Arno Press, 1955.

Murphy, Allan H. and Winkler, Robert L., "Probability Forecasting in Meteorology," *Journal of the American Statistical Association*, September 1984, *79*, 489–500.

Muth, John, "Optimal Properties of Exponentially Weighted Forecasts of Time Series with Permanent and Transitory Components," *Journal of the American Statistical Association*, June 1960, *55*, 299–306.

______, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

______, "Short Run Forecasts of Business Activity," paper presented at the joint Pittsburgh Meetings of the Eastern Economics Association–International Society for Inventory Research, March 1985.

Nerlove, Marc, "On the Optimality of Adaptive Forecasting," *Management Science*, January 1964, *10*, 207–24.

Prescott, Edward, "Adaptive Decision Rules for Macroeconomic Planning," *Western Economic Journal*, December 1971, *9*, 369–78.

______, "Should Control Theory be Used for Economic Stabilization?," in Karl Brunner and Alan Meltzer, eds., *Optimal Policies, Control Theory, and Technological Exports*, Vol. 7, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl., 1977, 13–38.

Sargent, Thomas J., "Beyond Demand and Supply Curves in Macroeconomics," *American Economic Review Proceedings*, May 1982, *72*, 382–89.

Shiller, Robert J., "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review," *Journal of Monetary Economics*, January 1978, *4*, 1–44.

Simon, Herbert G., "Rational Decision Making in Business Organizations," *American Economic Review*, September 1979, *69*, 493–513.

______, "Dynamic Programming Under Uncertainty with a Quadratic Criterion Function," *Econometrica*, January 1956, *24*, 74–81.

Theil, Henri, "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, April 1957, *55*, 346–49.

Tobin, James, *Asset Accumlation and Economic Activity: Reflections on Contemporary Macroeconomic Theory*, Chicago: University of Chicago Press, 1980.

Zellner, Arnold, "A Statistical Analysis of Provisional Estimates of Gross National Product and its Components, of Selected National Income Components, and Personal Saving," *Journal of the American Statistical Association*, March 1958, *53*, 54–65.

______, "Bayesian Econometrics," *Econometrica*, March 1985, *53*, 253–70.

Congressional Budget Office, "A Review of the Accuracy of Treasury Revenue Forecasts, 1963–1978," Staff Working Paper, Washington, February 1981.

Financial Panics, the Seasonality of the Nominal Interest Rate, and the Founding of the Fed

By JEFFREY A. MIRON*

After the founding of the Fed in 1914,[1] the frequency of financial panics and the size of the seasonal movements in nominal interest rates both declined substantially. Since the Fed was established in part to "furnish an elastic currency,"[2] it is natural to hypothesize that the Fed caused these changes in the behavior of financial markets. There were, however, a number of other major changes in the economy and in the financial system during this period including World War I, the shift from agriculture to manufacturing,[3] and the loosening of the gold standard.[4] Moreover, Robert Shiller (1980) has examined the effect of the Fed's founding on the seasonal in real interest rates and has concluded that the Fed's actions had little or no effect.

This paper investigates the relationship between financial panics, seasonal movements in nominal interest rates, and the open market operations of the Fed after 1914. The paper establishes that the Fed, by carrying out the seasonal open market policy that eliminated the seasonal in nominal interest rates, caused the decrease in the frequency of panics. Since seasonal movements are anticipated and financial panics are probably real events, the results show that an anticipated monetary policy had real effects on the economy.

The issue of whether anticipated monetary policy can affect real variables, which is at the heart of monetary economics, has received much recent attention following the well-known contributions by Robert Barro (1977, 1978). His results have been subjected to a barrage of critical review, much of it supporting his finding that only unanticipated changes in money have real effects (for example, Barro and Mark Rush, 1980; Robert Litterman and Lawrence Weiss, 1985; Robert Lucas, 1973; Shiller; Christopher Sims, 1980),[5] some of it arguing that the evidence rejects the neutrality of anticipated money (for example, Robert Gordon, 1982; Frederic Mishkin, 1982, 1983).[6] The generally inconclusive nature of the debate reflects the difficulty of determining whether policy caused or responded to changes in the economy and of distinguishing anticipated from unanticipated policy actions. Thomas Sargent (1976), when describing the possible observational equivalence of classical and nonclassical models, suggested that identification would be aided if it were possible to draw data from two different policy regimes.

*Department of Economics, University of Michigan, Ann Arbor, MI 48109. I thank Stanley Fischer, Larry Summers, Peter Temin, Olivier Blanchard, Milton Friedman, Steve Zeldes, Steve O'Connell, Sue Collins, Robert Clower, and two anonymous referees for helpful comments on earlier drafts of this paper.

[1] The Federal Reserve Act was passed by Congress on December 23, 1913. The Board of Governors took office and began planning the organization of the System on August 10, 1914. The twelve banks opened for business on November 16, 1914.

[2] This quote is from the preamble to the Federal Reserve Act.

[3] The share of agriculture in Gross Domestic Product fell from 24 percent in the period 1897–1901 to 12 percent in 1922. See *Historical Statistics of the United States,...* (1976, Series F125–129, p. 232).

[4] During World War I, several countries (including Great Britain) left the gold standard, so the United States was less affected by external conditions.

[5] The approach to testing neutrality in Barro and Rush is the same as in Barro (1977, 1978). Litterman-Weiss and Sims use vector autoregressive techniques and base their conclusions on the failure of money to be Granger causally prior for real income. Lucas shows in a cross section of countries that the variance of money shocks is negatively correlated with the variance of output movements.

[6] Barro (1978) introduced the use of cross-equation restrictions into this literature. This more powerful way of testing neutrality has been exploited extensively by Mishkin (1982, 1983), who has usually found that the data reject neutrality, contrary to the results of Barro.

The founding of the Fed and the subsequent smoothing of the seasonal pattern in nominal interest rates was a clear case of a change in regime, and the seasonal movements in the Fed's open market purchases were clearly an anticipated policy. It is the combination of these two things that I exploit in examining the neutrality of anticipated money.

I. A Model of the Banking System

This section presents a model of the banking system in which the magnitude of the seasonal movements in nominal interest rates is positively correlated with the frequency of financial panics.[7] The model shows that the Fed can reduce the frequency of panics by carrying out the seasonal open market policy that eliminates seasonal fluctuations from nominal rates. The model therefore suggests a channel through which anticipated monetary policy can have real effects.

The starting point is the Milton Friedman-Anna Schwartz (1963, pp. 50–53) textbook model of the money supply:

$$H = R + C \tag{1}$$

$$M = C + D \tag{2}$$

$$L = M - H \tag{3}$$

where H is high-powered money, M is money, C is currency, D is deposits, and L is loans. In this framework the money supply is determined by the interaction of the nonbank public (through the desired currency-deposit ratio), banks (through the desired reserve-deposit ratio), and the monetary authority (through high-powered money). The model presented here explicitly examines the bank's choice of reserve-deposit ratio and then makes standard assumptions about the remaining terms.

The banking system consists of a fixed number of identical banks, each of which is sufficiently small that it acts as a price taker. The representative bank holds two types of assets: reserves, R; loans, L. There is one type of liability: deposits, D. The bank accepts deposits infinitely elastically and pays out currency on demand.[8] The only decision it faces is what proportion of its assets to hold as reserves and what proportion as loans. The larger the proportion of loans, the greater the costs to the bank of managing its portfolio.

There are costs to the bank of holding a large proportion of its assets as loans because it can suffer unexpected deposit withdrawals. Under fractional reserve banking, a sufficiently large amount of withdrawals causes the bank to fail because some of its assets are tied up in loans and it takes time to convert these into cash. If the bank experiences withdrawals, therefore, it liquidates some of its loans to bolster its reserve position. This imposes costs since the bank accrues capital losses and/or incurs excess brokerage fees when it calls in loans unexpectedly.

The bank's cost function takes the form

$$c\left(\frac{R}{D}\right) = \frac{(W - E(W))^2}{2}\left(\left(\frac{R}{D}\right) - 1\right)^2, \tag{4}$$

where W is the amount of withdrawals that the bank experiences. Costs depend on the amount of unexpected withdrawals and on the ratio of reserves to deposits. They increase with the amount of unexpected withdrawals but decrease with the reserve-deposit ratio. The cost function described by (4) assumes that unexpected withdrawals and unexpected deposits have the same effect on costs. It also assumes that the distribution of withdrawals is independent of the level of deposits. Both of these assumptions are probably unrealistic, but they simplify the presentation of the results. The results do not depend on these two assumptions.

[7]For other recent examples of models of panics see John Bryant (1980), Douglas Diamond and Philip Dybvig (1983), and Gary Gorton (1982).

[8]In the pre-Fed period, demand and time deposits were much closer substitutes than they are today. Demand deposits sometimes paid interest, and time deposits could be transferred by check. The data do not distinguish between demand and time deposits (Friedman and Schwartz, p. 4). I assume here that all deposits are demand deposits and that they do not pay interest.

The bank's problem is

$$\max E(iL - c(R/D)) \tag{5}$$

subject to

$$R + L = D, \tag{6}$$

where i is the nominal interest rate. The solutions for R and L are

$$R^d = D(1 - (iD/s^2)) \tag{7}$$

$$L^s = D(iD/s^2), \tag{8}$$

where $s^2 = E(W - E(W))^2$. These solutions imply a desired loan-reserve ratio of

$$L^s/R^d = iD/(s^2 - iD). \tag{9}$$

When the interest rate or the level of deposits is high, the bank would like to hold a small proportion of its assets as reserves and a large proportion as loans. When the variance of withdrawals, s^2, is high, the bank wishes to hold a small proportion as loans and a large proportion as reserves.

To close the model I assume that the demand for loans is negatively related to the real interest rate and that the demand for deposits is interest inelastic:

$$L^d = P(Y - b(i - \pi^*)) \tag{10}$$

$$D^d = Pd, \tag{11}$$

where P is the price level and π^* is expected inflation. Real loan demand is negatively related to the real interest rate and positively related to Y, a measure of the real demand for credit in the economy. The demand for deposits, d, does not depend on the interest rate or on the state of the economy.

I assume for now that $P = 1$ and $\pi^* = 0$; this simplifies the presentation without affecting the results. I also assume that the level of high-powered money is fixed and independent of the behavior of the economy. The United States and its major trading partners were on the gold standard during the period 1890–1914, but there was sufficient sluggishness in gold flows so that the U.S. interest rate could move independently of the world rate in the short run.[9] In addition, as shown by Truman Clark (1983), there were seasonal movements in the world rate that corresponded closely to those in the United States.

Equations (7), (8), (10), and (11) jointly determine the equilibrium values of the endogenous variables in the model. The two exogenous variables Y and d parameterize the solutions. By noting how the solutions depend on these variables, we can see how they depend on external conditions. I interpret these external conditions as the effects of different seasons. Determining the sensitivity of the solutions to Y and d therefore tells us how equilibrium in financial markets depends on the seasonal movements in loan and deposit demand.

The equilibrium value of the interest rate is

$$i = Ys^2/(bs^2 + d^2). \tag{12}$$

The interest rate is high in seasons in which loan demand is high or deposit demand is low. When the variance of deposit withdrawals is high, the level of the interest rate is also high.

The equilibrium values for loans, reserves, and the loan-reserve ratio are

$$L = Yd^2/(bs^2 + d^2) \tag{13}$$

$$R = (bds^2 + d^3 - d^2Y)/(bs^2 + d^2) \tag{14}$$

$$L/R = Yd^2/(bds^2 + d^3 - d^2Y). \tag{15}$$

The quantity of loans is high when demand for them is high and when the deposits at banks are high; they are low when the variance of withdrawals is high. Reserves are low when loan demand is high and high when deposit demand is high. The ratio of loans to reserves increases with loan demand, decreases with deposit demand, and decreases with the variance of withdrawals.

The seasonal movements in Y and d also affect the distribution of costs of running the

[9] Friedman and Schwartz (pp. 89–90).

banking system:

$$(16)\quad c\left(\frac{R}{D}\right)=\frac{(W-E(W))^2}{2}\frac{d^2Y^2}{(bs^2+d^2)^2}.$$

These costs are high when d is low and when Y is high, given the distribution of W. That is, a withdrawal of a given size imposes higher costs on the banking system in periods when loan demand is high or deposit demand is low.

Panics can be thought of as periods when the costs of running the banking system are especially high. Since the distribution of costs shifts upward with the seasonal increases in loan demand and the seasonal decreases in deposit demand, the probability that costs exceed any given level is higher in seasons when loan demand is high or deposit demand is low. Thus panics are more likely to occur in these seasons.

This result, that panics are more likely to occur in seasons with high-loan demand or low-deposit demand, is the first key result provided by the model. The explanation is as follows. In some seasons there is an exogenous increase in loan demand that forces up nominal rates. Banks respond by loaning out a higher proportion of their reserves, which increases expected costs but also produces more revenue. The increase in loan-reserve ratios means a decrease in reserve-deposit ratios, so the distribution of costs, and thus the frequency of panics, is higher, even though the distribution of unexpected withdrawals is unchanged. It is the seasonal increase in loan demand and the resulting decrease in reserve-deposit ratios that causes an increased frequency of panics, not any change in the variance of deposit withdrawals.

I now examine the cost of running the banking system when the Fed intervenes by conducting open market operations. An open market purchase increases the supply of loans by an amount F. Assuming that this has no effect on P or π^*, the costs are

$$(17)\quad c\left(\frac{R}{D}\right)=\frac{(W-E(W))^2}{2}\frac{d^2(Y-F)^2}{(bs^2+d^2)^2}.$$

Costs decrease with F when $F<Y$. Since $c(R/D)$ is convex in F, the Fed can lower the average number of panics per year by conducting an open market policy that is seasonal and averages out to zero over the year. If open market operations affect P or π^*, then the derivation given above needs to be modified accordingly. In general, however, the Fed can still affect the behavior of nominal interest rates and therefore the frequency of panics.[10]

The conclusion that the Fed can reduce the frequency of panics by eliminating nominal interest rate seasonality is the second key result of the model. By supplying loans in periods when loan demand is high, the Fed accommodates the increase in demand and lessens the increase in the interest rate that would otherwise occur. This lessens the decrease in the reserve-deposit ratio and therefore the upward shift in the distribution of costs. Open market purchases in a season with high-loan demand thus decrease the probability of a panic.

The model presented above provides an explanation for the change in the behavior of panics and interest rates that occurred after 1914. The Fed accommodated the seasonal movements in loan demand, thereby smoothing the seasonal pattern in nominal interest rates. In response to the reduction in the seasonality of interest rates, banks reduced the seasonal variation in their desired reserve-deposit ratios, so in equilibrium these were smoother. Finally, the fact that reserve-deposit ratios were smoother meant that, on average, banks were less exposed to unexpected deposit withdrawals and so the frequency of panics fell.

The central implication of the hypothesis that the behavior of financial panics and interest rates changed after 1914 because of the Fed's seasonal open market operations is that there should have been seasonal movements in the amount of credit extended by the Fed. Additional implications are as follows: the total amount of credit outstanding in the economy should have become more seasonal after 1914; the loan-reserve ratio of banks should have become less seasonal; and

[10] I show this explicitly in my dissertation, ch. IV, which discusses seasonal movements in real rates.

the loans made by private banks should have become less seasonal.

Section II examines empirically the implications of the model and the hypotheses it suggests about the behavior of the Fed. These results come from a simple model, but they do not depend on the particular assumptions made in order to keep the analysis simple.[11] The model presented above is the most complicated one that can be tested empirically; it is not possible to test the additional implications of more complicated models because of data limitations.

II. The Evidence

A. *Historical Background: The National Banking System and the Founding of the Fed*

The period from 1863 through 1913 is known as the period of the National Banking System because the provisions of the National Banking Acts of 1863, 1864, and 1865 determined the banking and financial structure in several critical ways. The National Banking Acts were both a response to problems of the financial system that existed before the Civil War and a measure designed to raise revenue for the North during the war. The Acts successfully generated revenue and cured some prewar financial ills (notably the multiplicity of note issue). During the National Banking Period, however, those in academia, the banking community, and government still regarded the financial system as fundamentally flawed because of the "perverse elasticity of the money supply" and the high frequency of financial panics.

The term perverse elasticity of the money supply referred to the tendency of the money supply to contract in precisely those periods when it was "needed" most. This occurred in the spring and fall of each year when seasonal increases in loan and currency demand forced interest rates up and reserve-deposit ratios down. These seasonal movements in loan and currency demand were attributed mainly to the need for both currency and credit by the agricultural sector of the economy in the spring planting season and the fall crop-moving season, and to the need for currency and credit by the corporate sector for quarterly interest and dividend settlements. Additional currency was needed because the volume of transactions was higher in these periods. Credit demand was high because farmers borrowed to finance the planting and harvesting of the crops.[12]

The financial panics that occurred in this period were combinations of bank failures, bank runs, and stock market crashes. A typical panic began after an individual bank was hit by either an unexpectedly large deposit withdrawal or a large loan default. If the bank had a small amount of reserves, it would need to call in some of its loans. This might concern other banks enough so that they would call in some of their loans, many of which were in stock market call loans, and the cumulative effect of loan recall by many banks tended to depress the stock market. At the same time, the fact that banks were calling in loans caused the nonbank public to increase its desired currency-deposit ratio, and this could cause either individual bank failures or runs on many banks. Eventually the process either reversed itself or ended in a suspension of convertibility.[13]

There were, of course, differences in the dynamics of various panics. Some began in New York as the result of a large loan default at a New York bank and then were transmitted West as New York banks tried to acquire additional reserves from the country banks. Others started in the West when crop failures damaged the liquidity positions of country banks who then tried to recall

[11] Appendices A and B to ch. IV of my disseration show that the conclusions are still valid if one allows for a pyramided banking system, or for the general equilibrium interactions of the economy.

[12] E. W. Kemmerer (1910, pp. 223–24) mentions increased rail and barge activity during warm weather and holiday seasons as additional reasons for seasonal activity in the financial markets. A. Piatt Andrew (1906) discusses the influence of agriculture on economic activity during the pre-Fed period, and J. Laurence Laughlin (1912, pp. 309–42), discusses the seasonal cycle in general economic activity. See also O. M. W. Sprague (1910), C. A. E. Goodhart (1969), and John James (1978, pp. 127–37) for discussions of the seasonal flows within the country that accompanied the seasonal changes in interest rates and reserve positions of banks.

[13] Sprague (pp. 1–225).

balances from reserve cities. Nevertheless, the key element of a panic was the same in all of the major episodes. This key element was a generally increased demand for reserves that could not be satisfied for all parties simultaneously in the short run.

The likelihood that an event such as a large loan default would precipitate a panic depended on the initial position of the banking system. If such an event happened at a time when loan demand was high or deposit demand was low, so that the reserve-deposit ratios of banks were low, then the costs imposed by the loan default were higher. Since there were seasonal movements in loan and deposit demands that produced seasonal movements in reserve-deposit ratios, panics tended to occur in the fall and spring, when high-loan demand and low-deposit demand produced low reserve-deposit ratios. Thus the problems of perverse elasticity and the accompanying financial panics were partly a result of and coincided with the seasonal movements in asset demands.

The academics, bankers, and government officials of the time understood this phenomenon. J. Laurence Laughlin, a professor of economics at the University of Chicago, commented in detail on this relation between panics and seasonality in his 1912 treatise on reform of the banking system (pp. 309–42). Paul Warburg, a Wall Street banker who later served on the Federal Reserve Board, wrote in 1910 that "there can be no doubt whatever that the basis for healthy control by a central bank must exist in a country where regular seasonal requirements cause, with almost absolute regularity, acute increased demand for money and accommodation" (1930, p. 156). Leslie Shaw, Secretary of the Treasury from 1902 to 1906, actively attempted to accommodate the seasonal demands in financial markets, although the funds available to him were not sufficient to allow him to be successful.[14]

The panic of 1907 precipitated sufficient concern about panics and elasticity that Congress passed the Aldrich-Vreeland Act of 1908. This Act addressed the problems of the banking system by granting certain emergency powers to New York City banks and by creating the National Monetary Commission. This Commission was assigned to undertake a detailed study of the U.S. banking system. Its *Report*, published in 1910, contained in depth examinations of every aspect of banking theory and practice in the United States and abroad.

Two parts of the *Report* deserve particular notice. O. M. W. Sprague, a professor of economics at Harvard, wrote *History of Crises Under the National Banking System*. This book examined in detail the operation of the banking system during five of the worst financial crises (1873, 1884, 1890, 1893, and 1907). Sprague wrote that "with few exceptions all our crises, panics, and periods of less severe monetary stringency have occurred in the autumn" (p. 157). E. W. Kemmerer of Cornell contributed the volume *Seasonal Variations in the Relative Demands for Money and Capital in the United States*. He noted that "the evidence accordingly points to a tendency for the panics to occur during the seasons normally characterized by a stringent money market" (p. 232). Thus two parts of the *Report* mentioned explicitly the tendency for panics to occur in certain seasons of the year.

The Federal Reserve Act established the Federal Reserve System in 1913, three years after the publication of the Commission's *Report*. The preamble to the Act states that it is "an act to... furnish an elastic currency." It was to be expected, therefore, that the Fed would try to eliminate panics by accommodating the seasonal demands in financial markets.

B. *Evidence of the Changes in Financial Markets*

I now document the two facts cited in the introduction: the frequency of financial panics diminished after the founding of the Fed; and the size of the seasonal fluctuations in nominal interest rates diminished also.

Table 1 shows the starting dates of the financial panics that occurred during the period 1890–1908 according to Sprague and

[14]See Andrew (1907, p. 559), and Richard Timberlake (1978, p. 181). See Andrew (1907) also for an interesting analysis of Shaw's other activities and Timberlake (1963) for a critique of Andrew's analysis.

TABLE 1—STARTING DATES AND CLASSIFICATION OF FINANCIAL PANICS ACCORDING TO SPRAGUE AND KEMMERER

Classification	Year	Month
Sprague		
Financial Stringency	1890	August
Crisis	1893	May
Crisis	1907	October
Kemmerer		
Major Panics	1890	September
	1893	May
	1899	December
	1901	May
	1903	March
	1907	October
Minor Panics	1893	February
	1895	September
	1896	June
	1896	December
	1898	March
	1899	September
	1901	July
	1901	September
	1902	September
	1904	December
	1905	April
	1906	April
	1906	December
	1907	March
	1908	September

Sources: Sprague (pp. 1–225); Kemmerer (pp. 222–23).

Kemmerer.[15] Sprague classified periods of financial strain as either crises or "periods of financial stringency." A crisis was the more serious situation in his terminology and necessarily involved a suspension of convertibility of deposits into currency. There were three periods of serious strain according to Sprague, which amounted to one every six and one-third years. Kemmerer's classification system distinguished major from minor panics and included a larger number of episodes than Sprague's system. He determined the starting dates by reading the *Commercial and Financial Chronicle* and the *Financial Review*, two periodicals that were the *Business Weeks* of their day. Kemmerer found six major and fifteen minor panics during the 1890–1908 period. If only major panics are included, the frequency was slightly more than one every three years. Including minor panics raises the frequency to more than one per year.

Between 1915 and 1933, the banking system experienced financial panics only during the subperiod 1929–33.[16] There were several recessions during the subperiod 1915–28 (1918–19, 1920–21, 1923–24, 1926–27), one of which was quite severe (output fell 9 percent from 1920 to 1921). Nevertheless, until 1929 there were no financial disruptions of the types that occurred in the pre-Fed period, even during the recessions. The 1921 *Annual Report* of the Fed makes a point of noting that the financial market failures that had been symptomatic of earlier downturns did not occur during the 1920–21 recession.[17]

The question of whether the frequency of financial panics diminished after the founding of the Fed, therefore, consists of determining the probability that it would have taken fifteen years for the economy to experience its first panic after 1914 if in fact the tendency of the economy to panic had been unchanged. The appropriate data to use to estimate the frequency of panics during the pre-Fed period are Kemmerer's on the number of major panics; Sprague's definition of panics omits periods often cited elsewhere while Kemmerer's data on minor panics include periods that were not noted by many observers. Assuming that the distribution of panics was Bernoulli, Kemmerer's data provide an estimate of the probability of having a panic in a given year of .316. This implies that the probability of obtaining a sample of fourteen years with no panics was .005. The data therefore reject the hypothesis of no change in the frequency of panics at the 99 percent level of confidence.

Figure 1 shows the estimated seasonal pattern in the interest rate on stock market call loans for the periods 1890–1908 and 1919–

[15]Sprague and Kemmerer also give starting dates for panics for the 1873–89 period. Those data are qualitatively similar to those provided in Tables 2 and 3. I have not presented them because the interest rate and money market data are only available beginning in 1890.

[16]See Friedman and Schwartz (pp. 305, 308, 313, 324).

[17]Friedman and Shwartz (p. 235) and Phillip Cagan (1965, p. 225) also note the absence of financial crises during this recession.

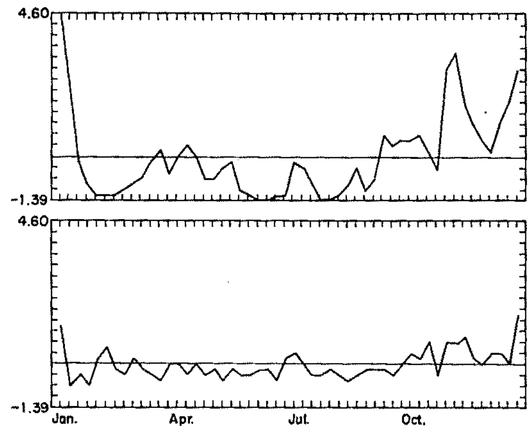


FIGURE 1. SEASONAL PATTERN IN NOMINAL INTEREST RATES, BEFORE (TOP) AND AFTER (BOTTOM) 1914

TABLE 2—TESTS OF THE NULL HYPOTHESIS OF NO SEASONAL FLUCTUATIONS IN FINANCIAL MARKET VARIABLES

Sample Period	Dependent Variable	F-Statistics	Significance Level
1890–1908	Nominal Interest Rate	1.68	.003
	Loans-Reserve Ratio	4.28	.000
	Loans	2.46	.000
	Reserves	4.90	.000
1919–28	Nominal Interest Rate	2.05	.000
	Loans-Reserve Ratio	4.90	.000
	Loans	3.65	.000
	Reserves	1.90	.000
1922–28	Reserve Credit	7.09	.000
	Total Credit	5.54	.000

28.[18] The patterns were calculated using weekly data by computing the unconditional mean in each week, after subtracting a trend. The top portion is for 1890–1908 and the bottom for 1919–28. Both are plotted in hundreds of basis points per year and show fifty-two coefficients. The patterns are statistically significant in each period, as reported in Table 2.

The size of the seasonal cycle clearly decreases from the earlier period to the latter.

[18] I present results for the 1890–1908 period in the text because that is the period for which Kemmerer and Sprague identified the dates of financial panics. Data for the entire 1890–1914 period confirm the results presented in the text.

TABLE 3—TESTS OF THE NULL HYPOTHESIS OF NO CHANGE IN THE PATTERN OF SEASONAL FLUCTUATIONS IN FINANCIAL MARKET VARIABLES

Sample Period	Dependent Variable	F-Statistics	Significance Level
1890–1908 vs.			
1919–28	Nominal Interest Rate	2.05	.000
	Loans-Reserve Ratio	4.90	.000
	Loans	3.65	.000
	Reserves	1.90	.000

The standard deviation of the seasonal cycle was 130 basis points before 1914, but only 46 basis points afterwards. The amplitude of the cycle dropped from 600 basis points before 1914 to 230 after. The change in the patterns is statistically significant, as shown in Table 3.

C. *Implications of the Model for the Pre-Fed Data*

The next step in the analysis is to confirm that the model presented in Section I is consistent with the data for the pre-Fed period. The central implication of that model is that the distribution of financial panics should have been seasonal, with periods of high frequency corresponding to periods of high interest rates. Table 4 summarizes the data in Table 1 on the starting dates of panics by showing the number of panics that began in each month of the year. It is clear that the distribution was seasonal and that the periods of high frequency were spring and fall. This result is confirmed by a χ^2 goodness-of-fit test of the null hypothesis that the number of panics that occurred in each month was the same; that is, a test of the null hypothesis of no seasonality. The calculated test statistic for the data in the table is 36.25 while the 99 percent critical value of the χ^2 statistic with 11 degrees of freedom is 24.73.

Figure 2 provides additional support for the model. It shows the estimated seasonal pattern in the call money loan rate (top) and the loan-reserve ratio (bottom) for the period 1890–1908. (The scale on the bottom is in percent and shows the percentage change in

TABLE 4—DISTRIBUTION OF FINANCIAL PANICS BY MONTH, 1890–1908

January	0
February	1
March	3
April	2
May	2
June	1
July	1
August	0
September	6
October	1
November	0
December	4

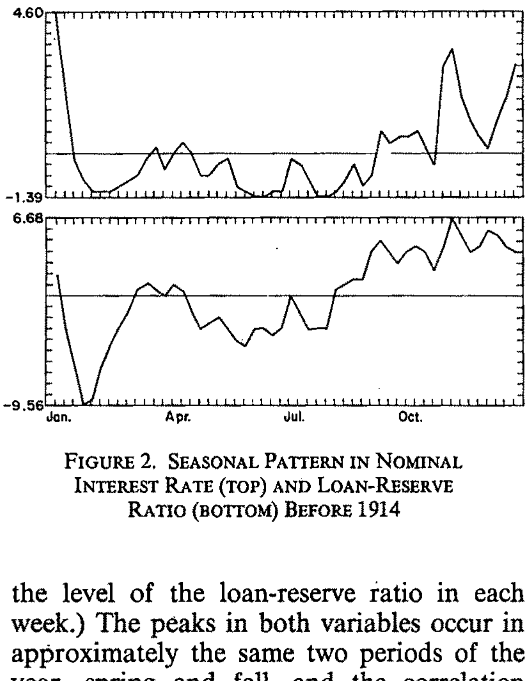


FIGURE 2. SEASONAL PATTERN IN NOMINAL INTEREST RATE (TOP) AND LOAN-RESERVE RATIO (BOTTOM) BEFORE 1914

the level of the loan-reserve ratio in each week.) The peaks in both variables occur in approximately the same two periods of the year, spring and fall, and the correlation between the estimated seasonal patterns is .63. Both the timing of the peaks, which coincides with that in financial panics, and the positive correlation between the seasonal patterns are results implied by the model.

Note that the pre-Fed seasonal movements in the loan-reserve ratio were large, with the standard deviation of the cycle being 3.5 percent and the amplitude 16.2 percent. In the post-World War II period, the elasticity of loan supply with respect to the interest rate has been small (Robert Rasche, 1972), and excess reserves have been kept near zero. The explanation for this change in behavior may be either the advent of FDIC or the smoother behavior of nominal interest rates. To the extent that the explanation is the lack of seasonal fluctuations in interest rates, this result also confirms the model.

D. *Implications of the Hypothesis that the Fed Caused the Changes in Financial Markets*

The hypothesis that the Fed caused the decrease in both the frequency of financial panics and the size of the seasonal movements in nominal interest rates implies that the actions of the Fed should have been seasonal, with the peaks of accommodation coming at those times of the year that had previously tended to be ones of financial stress. From 1915 to 1918, the Fed accommodated seasonal strain mainly by subsidizing loans for agricultural purposes, since the problems of financing World War I constrained its ability to conduct discretionary open market operations. Then, in 1918, it began to engage in significant seasonal open market operations.

The Fed established its loan subsidy program during the first full year of its existence, 1915. The program rediscounted bills backed by agricultural commodities at preferential rates in order to assure "that whatever funds might be necessary for the gradual and orderly marketing of the cotton crop" would be available. (The subsidy was not limited to bills backed by cotton, however.) The *Annual Reports* for the years 1916–18 all note that the program was working well and that the usual seasonal strain in financial markets had been avoided. The Fed discontinued the program in 1918 when it gained better control over its open market operations due to the end of the war.

Figure 3 shows the estimated seasonal pattern in Federal Reserve credit outstanding for the period 1922–28.[19] The periods of peaks in reserve credit outstanding coincide with the periods during which there were peaks in interest rates and loan-reserve ratios

[19] Weekly data are only available starting in 1922. Monthly data, which begin in August, 1917, confirm the results discussed here.

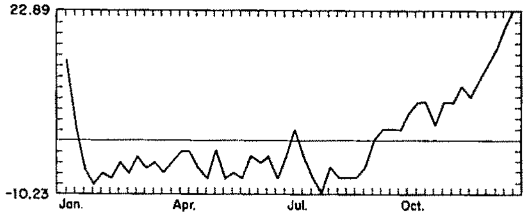


FIGURE 3. SEASONAL PATTERN IN FEDERAL RESERVE CREDIT OUTSTANDING

before the founding of the Fed, and the seasonal pattern is statistically significant (see Table 2). There was an increase of 32 percent in the level of reserve credit outstanding over the seasonal cycle, which amounted to roughly $400 million in a typical year. Since the total amount of loans outstanding at New York City banks was $6000 million, this was a substantial increase. It is clear that the actions of the Fed were seasonal and that they were likely to alleviate the seasonal strain that existed before 1914.

There are additional implications of the hypothesis of Fed responsibility that can be verified quantitatively. The seasonal variation in the total amount of credit outstanding should have increased after 1914 since, according to the hypothesis, the Fed's policy was one of subsidizing loan demand. Also, the seasonal variation in loan-reserve ratios should have decreased because this produced the seasonal variation in costs that the Fed wished to eliminate. Figure 4 shows the seasonal in loans by banks before 1914 and in the total amount of credit outstanding (banks and the Fed) after 1914. The total amount of credit outstanding became more seasonal after the founding of the Fed, with the standard deviation increasing from 1.4 to 1.8 percent and the amplitude rising from 4.8 to 7.7 percent. Figure 5 shows the estimated seasonal pattern in the loan-reserve ratio of banks before and after 1914. The figures show that the seasonal pattern diminished considerably, with the amplitude falling from 16.2 to 7.8 percent and the standard deviation dropping from 3.5 to 1.5 percent. Note in particular that those periods that showed the most significant seasonal strain before 1914 in all cases show almost none after 1914.

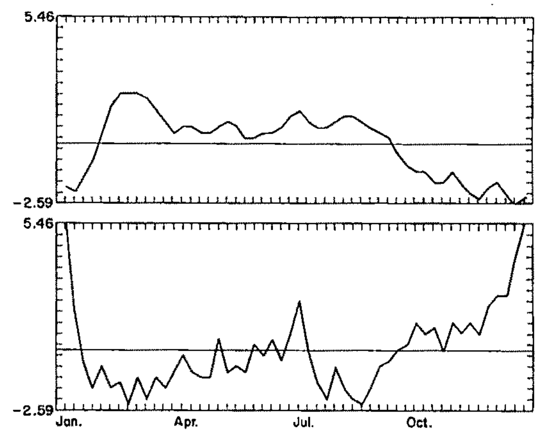


FIGURE 4. SEASONAL PATTERN IN TOTAL CREDIT OUTSTANDING BEFORE (TOP) AND AFTER (BOTTOM) 1914

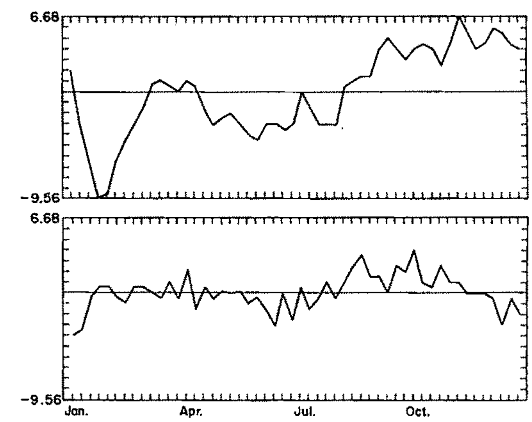


FIGURE 5. SEASONAL PATTERN IN LOAN-RESERVE RATIO BEFORE (TOP) AND AFTER (BOTTOM) 1914

The other implication of the hypothesis that the Fed caused the changes in financial markets is that the seasonal variation in loans made by private banks should have decreased. The seasonal pattern in loans by banks before and after 1914 is shown in Figure 6. The amount of seasonal variation diminishes, as implied by the hypothesis of Fed causation. The timing of the seasonal pattern also changes, however, and this is not implied by the hypothesis.

The explanation for the change in timing is probably that the pattern of deposits at banks changed after 1914 because the Treasury began keeping some of its deposits

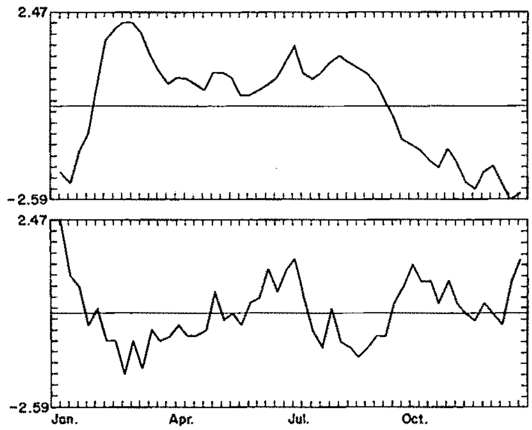


FIGURE 6. SEASONAL PATTERN IN LOANS BEFORE (TOP) AND AFTER (BOTTOM) 1914

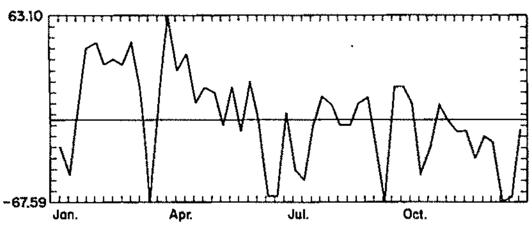


FIGURE 7. SEASONAL PATTERN IN TREASURY DEPOSITS AT THE FED

at the Fed. Before 1914, the Treasury kept its holdings of currency in either Treasury offices or private banks, and movements of cash between these two places affected the stock of high-powered money. After 1914 the Treasury also kept some of its deposits at the Fed. Since Treasury deposits at the Fed are not part of high-powered money, increases in Treasury deposits decrease high-powered money. If the Treasury kept at the Fed money that it had previously kept at private banks, the seasonal movements in the Treasury's holdings of cash would have produced different seasonal movements in high-powered money after 1914.

Figure 7 shows the seasonal pattern in Treasury deposits at the Fed for the period 1922:1–1928:12. There is a large peak at about the time of year (February/March) when the loan seasonal diminishes. Thus the change in the behavior of the Treasury may explain the change in the timing of the loan seasonal.

The quantitative evidence therefore establishes that the Fed caused the changes in the behavior of financial markets. Two quotes from the Fed's first *Annual Report* confirm this evidence. In discussing the proper role for monetary policy, the *Report* says,

> What is the proper place and function of the Federal Reserve Banks in our banking and credit system? On the one hand, it is represented that they are merely emergency banks to be resorted to for assistance only in time of abnormal stress; while on the other, it is claimed that they are in essence simply additional banks which should compete with the member banks, especially with those of the greatest power. The function of a reserve bank is not to be identified with either of these extremes. ...Its duty is not to await emergencies but by anticipation, to do what it can to prevent them. So also if, at any time, commerce, industry or agriculture are, in the opinion of the Federal Reserve Board, burdened unduly with excessive interest charges, it will be the clear and imperative duty of the Reserve Board acting through the discount rate and open market powers, to secure a wider diffusion of credit facilities at reasonable rate. ...The more complete adaptation of the credit mechanism and facilities of the country to the needs of industry, commerce, and agriculture—with all their seasonal fluctuations and contingencies—should be the constant aim of a Reserve Bank's management. [1914, p. 17]

Further, the *Report* states,

> It should not, however, be assumed that because a bank is a Reserve Bank its resources should be kept idle for use only in times of difficulty. ...Time and experience will show what the seasonal variates in the credit demand and facilities in each of the Reserve Banks of the several districts will be and when and to what extent a Reserve Bank may, without violating its special function as a guardian of banking reserves, engage in banking and credit operations. [1914, p. 18]

It is clear that the Fed considered the elimination of both seasonal strain and financial panics as essential parts of its function.

The statements of H. Parker Willis and Carter Glass provide additional support for this proposition. Willis, an economist at Columbia University, was an expert consultant to the House Banking and Currency Committee in 1912–13 while the Federal Reserve Act was being written, and he later became Secretary of the Federal Reserve Board. He wrote in 1915 that the potential benefits of the System were that "there will be no such wide fluctuations of interest rates ... from season to season as now exist ... and no necessity of emergency measures to safeguard the country from the possible results of financial panic or stringency" (p. 75). Carter Glass, who sponsored the Federal Reserve Act as a member of the House of Representatives in 1913, wrote in 1927 (p. 387) that two of the most important accomplishments of the System were the removal of panics and the elimination of seasonal interest rate fluctuations.

E. *Seasonality and the Financial Panics during the Great Depression*

The United States did not experience any financial panics from 1915 through 1928. There were, however, five financial panics during the period 1929–33, and these five were among the most severe in the country's history. Since the preceding sections have shown that the Fed successfully eliminated panics from 1915 to 1928, it is necessary to explain why panics recurred during the period 1929–33.

Figure 8 shows the actual level of Federal Reserve credit outstanding during the year 1929 as well as the level projected on the basis of the pattern that obtained from 1922 through 1928. The actual level is below the projected level starting in March. Further, the points at which the discrepancy increases correspond to periods that experienced peaks in loan demand during the 1890–1908 period. Figure 9 shows the estimated seasonal patterns in reserve credit outstanding for the two periods 1922–28 and 1929–33. The seasonal pattern is dampened in the later period, with the standard deviation of the cycle falling from 7.5 to 6.9 percent and the amplitude falling from 33.1 to 26.1 percent.

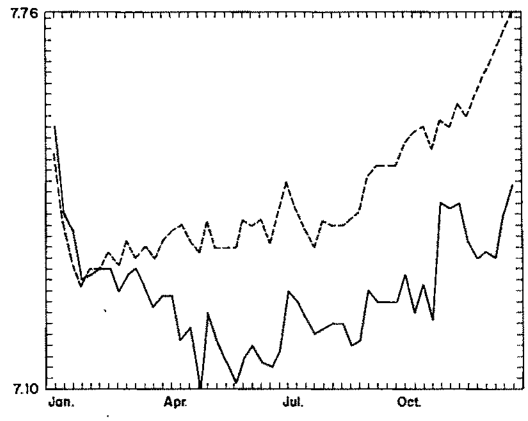


FIGURE 8. ACTUAL (SOLID LINE) AND FORECAST (BROKEN LINE) VALUE OF RESERVE CREDIT OUTSTANDING, 1929

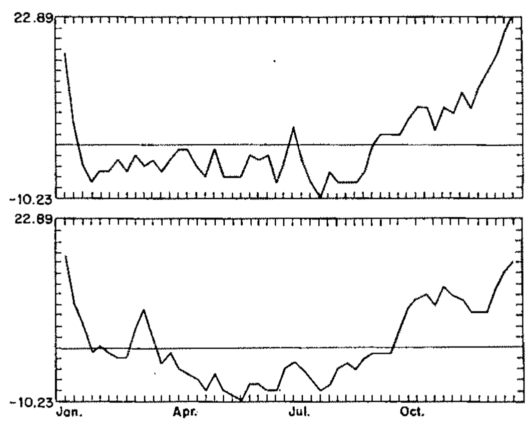


FIGURE 9. SEASONAL PATTERN IN RESERVE CREDIT OUTSTANDING, 1922–28 (TOP) AND 1929–33 (BOTTOM)

These results show that the Fed accommodated the seasonal demands in financial markets to a lesser extent during the 1929–33 period than it had previously. This means that the frequency of panics *should* have increased, as it did. It also means that the panics should have occurred in the spring and the fall, and this is exactly what happened: three of the panics were in the fall

(1929, 1930, 1932) while two were in the spring (1931, 1933).[20] Thus the recurrence of panics during this period corroborates the hypothesis that the Fed caused the reduction in the frequency of panics after 1914.

The decreased accommodation of the seasonals in asset demands was probably the result of a generally restrictive open market policy that the Fed initiated in late 1928.[21] During much of the 1920's, there was fear that speculation by participants in the stock market was "excessive," and those who objected to the speculation most encouraged the Fed to restrain the growth of credit, particularly loans by banks to stock market brokers. The officials at the Fed differed in their view of how much to restrain credit. On balance, however, they opposed restraining speculation so much that it might adversely affect general business activity.

This policy changed toward the end of 1928. Stock market speculation had been especially virulent, and the Fed responded with a strongly restrictive policy. The explanation for the change in policy is that Benjamin Strong, the governor of the New York Fed, died in October of 1928.[22] During the period 1915–28, Strong was a dominant force in the Federal Reserve System and in the entire financial community. In the words of his biographer, Lester Chandler, Strong was "one of the world's most influential leaders in the fields of money and finance. During the first fourteen turbulent, formative years of the Federal Reserve System, his was the greatest influence on American monetary and banking policies" (1958, p. 3). Strong intensely disliked stock market speculation, but was an outspoken critic of restraining speculation at the cost of causing a recession. His death allowed the balance of opinion at the Fed to shift toward greater restraint, and a highly restrictive policy resulted. One of the manifestations of this policy was the incomplete accommodation of the seasonal demands in financial markets.

F. *The Real Effects of Financial Panics*

The final issue discussed in this section is whether financial panics had real effects on the economy. It is not possible to test an explicit model of the real effects of panics because of data limitations.[23] It is possible, however, to demonstrate support for the proposition that panics had real effects if one is willing to make assumptions about what these effects might have been. I assume here that panics affected the distribution of output by decreasing the average level of real activity, increasing the variance of real activity, and increasing the length of business cycles.[24]

In the context of this paper there are two implications of the proposition that panics were real events. First, the distribution of output in the pre-Fed period should have been worse in panic years than in nonpanic years; second, the distribution of output post-Fed should have been better than that pre-Fed.

Table 5 shows the mean and variance of the rate of growth of annual real *GNP* for the period 1890–1908 and for this period minus the years in which panics occurred (Kemmerer's major panic definition). The average level of *GNP* growth is higher and the variance of real growth lower for nonpanic years, so these facts support the hypothesis that panics altered the distribution of output. They do not, of course, prove that hypothesis, since panics might be the result of negative output shocks. Nevertheless, the facts in the table lend plausibility to the proposition that panics changed the distribution of output during the pre-Fed period.

[20] See Friedman and Schwartz (pp. 305, 308, 313, 324).

[21] See Paul Trescott (1982) for a more detailed examination of this aspect of Fed policy.

[22] Friedman and Schwartz (pp. 411–19 and pp. 692–93) discuss in detail the effects of Strong's death on the power structure of the Fed.

[23] The data in Gordon are interpolations.

[24] See Ben Bernanke (1983) for an analysis of the effects of the financial crises during the Great Depression, Cagan for a discussion of the real effects of panics during the pre-Fed period, and Gorton (1983) for work on the general relation between panics and business cycles.

TABLE 5—THE EFFECTS OF PANICS ON REAL *GNP* GROWTH (Shown in Percent)

	Mean	Standard Deviation
1890–1908	3.75	5.83
1890–1908[a]	6.82	5.46

Notes: The definition of panics is Kemmerer's major panics.
[a]1890–1908 minus the years with panics and the years immediately following panics.

TABLE 6—THE EFFECT OF THE ELIMINATION OF PANICS ON THE LENGTH OF BUSINESS CYCLES

Peak	Trough	Length
July 1890	May 1891	11
January 1893	June 1894	18
December 1895	June 1897	19
June 1899	December 1900	19
September 1902	August 1904	24
May 1907	June 1908	14
August 1918	March 1919	9
January 1920	July 1921	19
May 1923	July 1924	15
October 1926	November 1927	14

Average for Pre-Fed Period = 17.5 months
Average for Post-Fed Period = 14.25 months

Source: Citibank Economic Database.

Table 6 compares the length of business cycles before and after the founding of the Fed. The measure of the length of a cycle is the number of months from peak to trough according to the dating of the National Bureau of Economic Research. I use this measure of the distribution of output rather than real *GNP* growth because the data on real *GNP* are not available on a consistent basis for the entire 1890–1928 period.

The data in the table show that the length of a typical recession fell after the founding of the Fed. The average length during the period 1890–1908 was 17.5 months while during the period 1919–28 it was only 14.25 months. More impressively, three of the four post-Fed recessions were as short or shorter than four of the six pre-Fed recessions. There are naturally many possible explanations for these facts, but they do provide basic support for the proposition that eliminating panics had real effects on the economy.[25]

III. Conclusions

It is well known that it is difficult to draw conclusions about the operation of monetary policy when data are drawn from a single policy regime. The founding of the Fed and the subsequent smoothing of the seasonal pattern in nominal interest rates constituted a clear shift in regime. In this paper I have exploited that change in regime, along with the fact that seasonal movements are anticipated, to reach conclusions about the effects of monetary policy, namely, that anticipated open market operations by the Fed probably had real effects.

The conclusion of this paper clearly raises the question of what current monetary policy should do about nominal interest rate seasonals. During the post-World War II period there has been substantial seasonality in both open market operations and the stock of money, but virtually no seasonality in nominal rates. This suggests that the Fed has accommodated seasonal fluctuations in asset demands since World War II in much the same way that it did following its inception in 1914. It also suggests that if the Fed produced a nonseasonal money stock, as Friedman (1959, 1982) has suggested it should, there would be a return of interest rate seasonals.

The return of nominal interest rate seasonals, however, would probably not cause a return of financial panics. Congress imposed deposit insurance on the banking system in 1934, and the presence of deposit insurance probably eliminates panics independently of the Fed's seasonal policy. Thus the analysis

[25]In ch. IV of my dissertation, I also discuss three alternative explanations for the disappearence of interest rate seasonality after 1914. In particular, I address Clark's claim that the reduction in seasonality could not have been due to the Fed, since seasonality in nominal rates disappeared worldwide after 1914. I show that this was probably the result of attempts by many central banks to eliminate interest rate seasonals. It therefore supports rather than contradicts my explanation of the U.S. experience.

above does not necessarily imply that continued elimination of interest rate seasonals is desirable. The analysis does show that an important aspect of Fed policy is its seasonal behavior, and it demonstrates that this aspect of policy can have substantial real effects on the economy.

REFERENCES

Andrew, A. Piatt, "The Influence of Crops Upon Business in America," *Quarterly Journal of Economics*, May 1906, *20*, 323–53.

______, "The Treasury and the Banks Under Secretary Shaw," *Quarterly Journal of Economics*, August 1907, *21*, 519–68.

Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

______, "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August 1978, *86*, 549–80.

______ **and Rush, Mark,** "Unanticipated Money and Economic Activity," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980, 27–74.

Bernanke, Ben S., "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, June 1983, *73*, 257–76.

Bryant, John, "A Model of Reserves, Bank Runs, and Deposit Insurance," *Journal of Banking and Finance*, December 1980, *4*, 335–44.

Cagan, Phillip, *Determinants and Effects of Changes in the Stock of Money, 1875–1960*, New York: Columbia University Press, 1965.

Chandler, Lester, *Benjamin Strong, Central Banker*, Washington: The Brookings Institution, 1958.

Clark, Truman, "Interest Rate Seasonals and the Federal Reserve System," unpublished manuscript, University of Southern California School of Business Administration, 1983.

Diamond, Douglas W. and Dybvig, Philip H., "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, June 1983, *91*, 401–19.

Friedman, Milton, *A Program for Monetary Stability*, New York: Fordham University Press, 1959.

______, "Monetary Policy: Theory and Practice," *Journal of Money, Credit and Banking*, February 1982, *14*, 98–118.

Friedman, Milton and Schwartz, Anna J., *A Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press, 1963.

Glass, Carter, *An Adventure in Constructive Finance*, New York: Doubleday, Page, and Co., 1927.

Goodhart, C. A. E., *The New York Money Market and the Finance of Trade, 1900–1913*, Cambridge: Harvard University Press, 1969.

Gordon, Robert J., "Price Inertia and Policy Ineffectiveness in the United States, 1890–1980," *Journal of Political Economy*, December 1982, *90*, 1087–117.

Gorton, Gary, "Demand Deposits and Banking Panics," manuscript, Wharton School, University of Pennsylvania, 1982.

______, "Banking Panics and Business Cycles," manuscript, Wharton School, University of Pennsylvania, 1983.

James, John A., *Money and Capital in Postbellum America*, Princeton: Princeton University Press, 1978.

Kemmerer, E. W., *Seasonal Variations in the Relative Demand for Money and Capital in the United States*, National Monetary Commission, S.Doc.588, 61st Cong., 2d session, 1910.

Laughlin, J. Laurence, *Banking Reform*, Chicago: National Citizens League for the Promotion of a Sound Banking System, 1912.

Litterman, Robert B. and Weiss, Laurence, "Money, Real Interest Rates, and Output: A Reinterpretation of Postwar United States Data," *Econometrica*, January 1985, *53*, 129–56.

Lucas, Robert E., Jr., "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

Miron, Jeffrey A., "The Economics of Seasonal Time Series," unpublished doctoral

dissertation, MIT, 1984.

Mishkin, Frederic S., "Does Anticipated Policy Matter? An Econometric Investigation," *Journal of Political Economy*, February 1982, *90*, 22–51.

______, "Does Anticipated Aggregate Demand Policy Matter? Further Econometric Results," *American Economic Review*, September 1983, *72*, 788–802.

Rasche, Robert, "A Review of Empirical Studies of the Money Supply Mechanism," *Federal Reserve Bank of St. Louis Review*, July 1972, *54*, 11–19.

Sargent, Thomas, "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *Journal of Political Economy*, July 1976, *84*, 631–40.

Shiller, Robert, "Can the Fed Control Real Interest Rates?," in Stanley Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980, 117–68.

Sims, Christopher, "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review Proceedings*, May 1980, *70*, 250–57.

Sprague, O. M. W., *History of Crises Under the National Banking System*, National Monetary Commission, S.Doc.538, 61st Cong., 2d session, 1910.

Timberlake, Richard H., "Mr. Shaw and His Critics: Monetary Policy in the Golden Era Reviewed," *Quarterly Journal of Economics*, February 1963, *77*, 41–54.

______, *The Origins of Central Banking in the United States*, Cambridge: Harvard University Press, 1978.

Trescott, Paul B., "Federal Reserve Policy in the Great Contraction: A Counterfactual Assessment," *Explorations in Economic History*, July 1982, *19*, 211–20.

Warburg, Paul M., *The Federal Reserve System: Its Origins and Growth*, Vol. II, New York: Macmillan, 1930.

Willis, H. Parker, *The Federal Reserve: A Study of the Banking System of the United States*, New York: Doubleday, Page and Co., 1915.

Board of Governors of the Federal Reserve System, *Annual Report*, Washington: USGPO, 1914–33.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970*, New York: Basic Books, 1976.

Productivity, *R&D*, and Basic Research at the Firm Level in the 1970's

By ZVI GRILICHES*

This paper reports new results on the relationship of research and development (*R&D*) expenditures, especially expenditures on basic research, to productivity growth in U.S. manufacturing firms during the 1970's. It is based on a unique data set, the National Science Foundation (NSF) *R&D*-Census match, containing information on *R&D* expenditures, sales, employment, and other detail for approximately 1000 largest manufacturing firms from 1957 through 1977. It updates my earlier work (1980) on the precursor of this data set, replicates some of Edwin Mansfield's (1980) work on the contribution of basic research to productivity growth using a larger, more recent, and more representative sample of firms, and complements similar work by myself and J. Mairesse (1983,1984) based on a publicly accessible but more limited data set.

Two topics are explored in some detail: 1) Is there any evidence of a decline in the returns to industrial *R&D* expenditures, a decline in their "fecundity" in the 1970's as compared to earlier time periods? 2) Is there evidence that basic research is a relatively more important component of *R&D* and that there may have been an underinvestment in this component?

A few background facts are worth stressing at this point. In the United States, total *R&D* expenditures in industry peaked (in real terms) around 1968, dropped slightly in the early 1970's and recovered somewhat in the late 1970's. Relative to total sales, *R&D* expenditures in industry declined from 4.2 percent in 1968 to a trough of 2.6 percent in 1979 and then recovered to 3.7 percent by 1982. This pattern masks a strong divergence between the trends in federally and privately supported industrial *R&D*. Federally supported *R&D* fell from 2.1 percent of manufacturing sales in 1967 to 0.7 percent in 1979 and has only recently begun to recover, while company-financed *R&D* stayed essentially constant (relative to industry sales) with almost all of the fluctuation coming from the decline in federal support (NSF, 1983; 1984). During the same period, the economy experienced one of the sharpest and prolonged recessions of the postwar period and a large and pervasive productivity slowdown. Hardest hit were the primary metals, motor vehicles, and other heavy, energy-related industries. On the whole, these were the less *R&D* intensive industries, resulting in a largely accidental correlation between *R&D* intensiveness and the productivity slowdown. (See my 1980 papers and my article with F. Lichtenberg, 1984, for more discussion of these issues.)

The remainder of the paper is organized as follows. First, I describe the data set with its advantages and limitations and present some overall comparative statistics. Second, I outline briefly the framework that underlies the computations to be performed. The results are presented and discussed, and the paper closes with some conclusions, caveats, and suggestions for further research.

I. Previous Work and the Current Data Set

The current project is an extension of work originally begun in the mid-1960's. That work was based on the matching of *R&D* data

*Harvard University, 125 Littauer Center, Cambridge, MA 02138, and the National Bureau of Economic Research. I am indebted to NSF Grant no. SES-82-08006 for the support of this work, to Douglas Dobas and Bronwyn H. Hall for making the data gathering effort possible, and to David Body for research assistance. I have also benefited from comments of seminar participants at the NBER and Yale University.

collected on behalf of the NSF by the Bureau of the Census during 1957–65 with additional company data from the 1958 and 1963 Census of Manufactures and Enterprise Statistics. The universe consisted of large (1000 or more employees) U.S. manufacturing companies performing *R&D*. The final sample of 883 of such companies accounted for over 90 percent of total sales and *R&D* expenditures of all firms in this universe.

The main finding of that work (see my 1980a paper) was a rather consistent and positive relationship between various measures of company productivity and its investments in research and development. The Cobb-Douglas-type production functions, estimated on both levels (1963) and rates of growth (1957–65) yielded an elasticity of output with respect to *R&D* investments of about .07 and an implied average gross excess rate of return of 27 percent (as of 1963), a significantly lower rate of return to federally financed *R&D* expenditures, and no clear evidence of significant scale effects either in *R&D* investment policies or the returns from it.

In trying to extend the earlier study to the more recent time period, it became clear that the earlier work could not be simply updated because much of the earlier data was lost and a new data set had to be created instead. The basic objective was to create a matched body of data on most of the large *R&D* performing corporations in the United States, making it possible to analyze both the determinants and consequences of *R&D* spending *over time*. For this purpose a time-series record has been created for each company consisting of the major variables in the annual *R&D* survey for each of the years 1957–77, supplementary *R&D* information for selected years (1962, 1967, 1972, and 1975), data from the Enterprise Statistics (i.e., company level questionnaires) for 1967, 1972, and 1977, and a few additional items from the Census of Manufactures establishment record summaries for 1967 and 1972. The data set began with all the "certainty" companies in the NSF *R&D* survey as they existed in 1972. There were approximately 1100 such companies, but a "complete" record is available only for a much smaller number.[1]

Table 1 lists the sample size, means, and standard deviation for the major variables as of 1972 and their growth rates from 1966 to 1977. It is intended to describe three aspects of these data: 1) the general characteristics (means and standard deviations) of the sample as of 1972; 2) average rates of growth of the major variables of interest during the 1967–77 period; and 3) how these measures change when the sample is changed to select observations according to the availability of the requisite information.

Turning to the last topic first, note that we tend to lose smaller and more *R&D* intensive firms as the sample gets more restrictive. The first column of Table 1 corresponds to the most liberal criterion: a firm had to exist in 1972 and report positive *R&D*. Column 2 requires both the ability to compute a growth rate for the 1967–77 period (i.e., at least five good time-series observations) and a successful match to the 1972 Census of Enterprise data (NCK–1). In column 3, I add the requirement of a successful match to the 1977 Census data, while in column 4 the subsample is based on a match with the 1967 and 1972 Census data instead. The major differences occur in the transition from column 1 to column 2 where trying to match to the Census we lose a relatively large number of smaller firms for which there are still data in the *R&D* survey files. The firms that can be also found in the 1977 Census are slightly larger and have had a somewhat higher rate of growth in employment, *R&D*, and productivity. The firms that also existed in 1967

[1]The universe of this data match consists of all "certainty" cases in the 1972 *R&D* survey; i.e., the basic definition is the population of companies as they existed in 1972 (as against 1962 in the earlier study) and the requirement of "certainty" assures that the Census Bureau tried to collect consistent data for these firms for more than one year. The "certainty" cases correspond closely to the earlier restriction to companies with 1000 or more employees, though it is a bit more inclusive. See my paper with Bronwyn Hall (1982) and Hall (1984) for more detail on sample definition and variable construction.

TABLE 1—MAJOR VARIABLES IN 1972 AND 1966-77 GROWTH RATES BY SUBSAMPLE: MEANS AND STANDARD DEVIATIONS[a]

Variable	Data Set, Selection Criteria, and Sample Size			
	1972 *R&D* Survey Universe ($N=1105$) (1)	1966–77 Growth Rates Computable and Matched to 1972 Census ($N=652$) (2)	(2) and Matched to 1977 Census ($N=491$) (3)	(2) and Matched to 1967 Census ($N=386$) (4)
A. Levels in 1972				
Sales in Million Dollars	146	205	223	236
	(1.61)	(1.43)	(1.40)	(1.44)
Total Employment	4038	5570	6212	6698
	(1.48)	(1.27)	(1.30)	(1.31)
R&D Scientists and Engineers	89	74	82	106
	(1.66)	(1.70)	(1.71)	(1.72)
R&D in Million Dollars	2.3	3.0	3.4	4.3
	(1.74)	(1.78)	(1.77)	(1.83)
R&D to Sales Ratio (*RS*)	.051	.033	.032	.035
	(.131)	(.064)	(.051)	(.048)
Company *R&D*/Sales Ratio (*CRS*)	.028	.022	.023	.025
	(.069)	(.026)	(.026)	(.026)
Basic to Total *R&D* Ratio (*BR*)	.025	.026	.026	.027
	(.074)	(.071)	(.075)	(.073)
Value-Added, Million Dollars		100	113	121
		(1.32)	(1.31)	(1.34)
Gross Fixed Assets Million Dollars		115	124	147
		(1.67)	(1.59)	(1.65)
B. Growth Rates 1966–77				
Employment Growth		.012	.015	.006
		(.046)	(.041)	(.040)
Partial Productivity Growth (*BPT*)		.025	0.26	.025
		(.036)	(.034)	(.035)
Total *R&D* Growth, Deflated (*BTRD*)		−.001	.003	−.007
		(.079)	(.074)	(.070)
Scientists and Engineers Growth		.008	.012	.004
		(.087)	(.084)	(.078)
Company *R&D* Growth, Deflated (*BCRD*)		.004	.008	−.000
		(.081)	(.076)	(.071)

Notes: Col. 1: "Certainty" firms in the NSF *R&D* Survey with positive *R&D* in 1972; Col. 2: Growth rates for 1966–77 computable (at least 5 years of good data on sales, employment, and *R&D*) and a successful match to the 1972 Enterprise Census (NCK–1); Col. 3: (2) and a successful match to the 1977 Census (NCK–1); Col. 4: (2) and a successful match to the 1967 Census and growth rates computable for 1957–65; Partial productivity growth = deflated sales growth − (share of labor compensation in total sales) × growth in employment; Sales deflated by NIPA based output price indexes at the 2–3 digit SIC level. *R&D* deflator based on the methodology suggested by Jaffe (NSF), from my 1984 comment.

[a]Geometric means and standard deviations (shown in parenthesis) of the logarithms (approximate coefficient of variation) except for growth rates or ratios.

are even larger but have on average grown somewhat more slowly than those that existed in the 1972–77 period. If we look at two of the major variables of interest, partial productivity growth and the ratio of basic to total *R&D*, there is almost no difference in their means across the relevant columns (2, 3, and 4), and hence it is unlikely that subsequent conclusions will be subject to a serious sample selection bias. I will, therefore, ignore this topic here.

Looking at the levels of the variables in 1972, we see that the average firm in the sample is quite large (5000+ employees),

employs close to one hundred $R\&D$ scientists and engineers, and is making only a relatively modest investment of its own money (about 2.5 percent of sales) in $R\&D$, with very little of that, less than 3 percent, being devoted to basic $R\&D$. This picture is somewhat misleading, however. The actual distribution of firms is quite skewed, with a small number of larger firms spending much larger amounts on both total and basic $R\&D$. Looking at growth rates one can observe that on average these firms grew only moderately during this period: about 1 percent per year in total employment, about 2.5 percent per year in partial productivity, and almost zero growth in deflated $R\&D$ expenditures (though a slightly positive rate of growth in the number of $R\&D$ scientists and engineers). Here again, while on average there is little movement, there is a great deal of variability at the individual firm level. The standard deviations of the rates of growth of partial productivity and total $R\&D$ are 3.5 and 8 percent per annum, respectively, with many firms growing much faster (and also much slower) than the average.

Looking at some of the $R\&D$ ratios over time, not reported in Table 1, one cannot see any significant decline in the rate of private investment in $R\&D$. While the total $R\&D$ to sales ratio falls from .042 in 1962 to .035 in 1972 and again from .032 in 1972 to .029 in 1977 for firms in subsamples 4 and 3, respectively, the company-financed $R\&D$ to sales ratios (CRS) are essentially unchanged (.025 in 1962 and 1972 in subsample 4 and .023 in 1972 and 1977 in subsample 3). On the other hand, while the basic research ratio (BR) fell only modestly from .033 to .031 between 1962 and 1972, and from .027 to .023 between 1972 and 1977, coupled with the decline in the overall total $R\&D$ to sales ratio, this implies about a 40 percent reduction in the relative intensity of industrial investment in basic research, relative to industry sales. Almost all of this decline came from the overall decline in federally financed $R\&D$ which declined from about 55 percent of total $R\&D$ in industry in 1965 to about 35 percent in 1982. The federal government financed about 32 percent of all basic research in industry in 1967 but only 19 percent in 1982 (see NSF, 83–302 and 84–311). The reduction was so steep that basic research in industry declined not only relatively (to sales) but also absolutely, from a peak of \$813 million in 1966 (in 1977 dollars) to a trough of \$581 million in 1975 and did not surpass the 1960's levels until the early 1980's. How one interprets the consequences of such a decline depends on one's view of the relative productivity of governmentally financed $R\&D$ expenditures in industry, a topic I will be exploring below.

II. The Analytical Framework and Econometric Results

The work reported here focuses primarily on the analysis of productivity growth for these companies, using a rather simple Cobb-Douglas production function approach:

$$(1) \qquad Q_t = Ae^{\lambda t}K_t^{\alpha}C_t^{\beta}L_t^{1-\beta},$$

where Q is output (sales, or value-added), C and L are measures of capital and labor input, respectively, $K=\Sigma_i w_i R_{t-i}$ is a measure of the accumulated and still productive research capital ("knowledge"), R_t measures the real (deflated) gross investment in research in period t, and the w_i's connect the levels of past research to the current state of knowledge. In addition, λ measures the rate of disembodied "external" technical change (where t is time in years), A is a constant, and constant returns to scale have been assumed with respect to the conventional inputs (C and L).

A number of serious difficulties arise when one turns to the operational construction of the various variables (see my 1979 article for more detailed discussion). Perhaps the two most important problems are the measurement of output (Q) in a research-intensive industry (where quality changes may be rampant), and the construction of the unobservable research capital measure (K). Turning to the second problem first, note that $K_t=\Sigma w_i R_{t-i}$ can be thought of as a measure of the distributed lag effect of past research investments on productivity. There are at least three forces at work here: the lag be-

tween investment in research and the actual invention of a new technique or product; the lag between invention and the development and complete market acceptance of the new product; and its disappearance from the currently utilized stock of knowledge due to changes in external circumstances and the development of superior techniques or products by competitors (depreciation and obsolescence). There is some scattered evidence, based largely on questionnaire studies that such lags are rather short in industry, where most of research expenditures are spent on development and applied topics, and where the private returns from *R&D* become obsolete much faster due to the erosion of a firm's specific monopoly position (Ariel Pakes and M. Schankerman, 1984).

While my models are written as if the main point of research expenditures is to increase the physical productivity of the firm's production process, most of the actual research in industry is devoted to the development of new products or processes to be sold and used outside the firm in question. Assuming that, on average, the outside world pays for these products what they are worth to it, using sales or value-added as the dependent variable does, in fact, capture the private returns to such research endeavours. However, the observed private returns may underestimate the social returns because, given the competitive structure of the particular industry, the firm is unlikely to appropriate all of these returns. On the other hand, part of the increase in the revenues of a particular firm may come at the expense of other firms, or from changes in the market power induced by the success of its research program. I cannot say much about the net impact of such forces on the basis of the data at hand. This would require a detailed comparison of the individual firm results with estimates based on industry and economy-wide returns to research, a topic beyond the scope of this project. But since expected private returns are a determinant of private investment flows into this activity, they are of some interest even if one cannot answer the social returns question unequivocally.

This framework can be extended to ask whether different types of *R&D* (private vs. federal, or basic vs. applied) are equally "potent" in generating productivity growth. One way of answering this question is to look at the "mix" of *R&D* expenditures and ask if it matters for the question at hand. Let there be two types of *R&D* expenditures, R_1 and R_2, and let us assume that the overall analysis is in terms of the logarithm of total *R&D* expenditures but that we believe that R_2 should have been weighted more, given a δ premium (or discount). That is, the right variable is

$$(2)\qquad R^* = R_1 + (1+\delta)R_2 = R(1+\delta s),$$

where $s = R_2/R$ is the "share" of R_2 in total $R = R_1 + R_2$. Then the α log R^* term can be approximated by $\alpha \log R^* \simeq \alpha \log R + \alpha\delta s$. The sign and significance of the mix term s will give us some clue about the size and magnitude of the δ term.

A similar argument can be made also in the context of a growth-rate formulation. Let lower case letters denote growth rates. Then $r = (1-s)r_1 + sr_2$ while $r^* = (1-s)r_1 + (1+\delta)sr_2$. If, as is mostly the case in our data, the growth rates of r_1 and r_2 are roughly equal, then $r^* = r(1+\delta s)$, and again, the coefficient of the mix term s provides us with some information about the "premium" or "discount" on R_2 since αr^* can be approximated by

$$(3)\qquad \alpha r^* \simeq (\alpha + \delta\bar{s})r + (\alpha\bar{r}\delta)s.$$

Given the peculiarities of my data set—its unbalanced nature (many missing observations towards the beginning and end of the period), the availability of capital and value-added only for Census years, the desire to preserve comparability with the earlier study, and the difficulty of doing elaborate programming inside the Census Bureau, I focus primarily on two major dimensions of the data: levels (in 1967, 1972, and 1977) and growth rates, and eschew any attempt at a complete annual data analysis. The annual data are summarized by computing average growth rates for two subperiods 1957–65 (corresponding to the earlier study period) and 1966–77, based on regressions of the logarithms of the relevant variables on time

trends (solving thereby the missing years problem within each of these subperiods).

In implementing such a framework of analysis one has to deal with several serious data problems: missing data, erroneous data and possible erroneous matches, and mergers. Except for *R&D* data, no special effort was made to replace missing values by various imputation procedures. It was my notion that the basic data set represents what the Census did collect, what we actually know, and that any imputation procedure should be done only in the context of a particular research project where its implications for the final analysis could be interpreted. As far as the *R&D* data are concerned, the Census used the shuttle nature of the original questionnaires to fill in many of the original blanks. To the extent that there remain missing values which are not due to the fact that the whole company is missing before or after some date, they were interpolated on the basis of the estimated growth rates (which require at least five good data points within each subperiod). For other variables, missing values were not imputed. It was not possible, within the constraints of this project, to develop optimal imputation procedures. This would have required several repeated passes at the original numbers. Instead, the analysis is based either on reduced "clean" samples or on "pairwise present" correlation coefficient matrices.

From an econometric point of view, we have to deal with the problem of firm effects (or firm-specific left-out variables) and the possibility that the relationships being estimated may not stay constant either across firms or across time. The first is handled by analyzing first differences or growth rates, transformations that eliminate any unchanging effects from the data. The second problem, the problem of differences across firms, is handled in part by calculating a measure of "partial" productivity growth $[BPT = y - (1-\hat{\beta})l]$, using individual firm data on the share of labor in total costs. One can also estimate separate and different parameters for the various industry groupings and include some of the other variables available in the record which might distinguish one firm's environment and response pattern from another's (such as its specialization ratio, size, or vertical integration). The main hypothesis under investigation, that the returns to *R&D* investments may have declined over time, is tested both by comparing estimates based on the more recent data with the earlier results, and by allowing and testing for systematic changes in the estimated relationships between the three available cross sections.

Let us look now at the first set of substantive results. Table 2 reports the results of estimating cross-sectional production functions (equation (1)) separately for each Census year, adding to the standard capital and labor variables a measure of total *R&D* capital accumulated by the firm and two *R&D* mix variables: the fraction of total *R&D* that was spent on basic research and the fraction of accumulated *R&D* that had been financed privately. All the reported estimates allow for 18 to 20 (depending on the subsample) separate industry intercepts. Columns 1 and 3 report estimates that are based on the same number of firms and use the same dependent variables, differing only by the year of observation. Column 2 presents additional estimates for 1972 based on different sample and dependent variable definitions with the main intent being to show that the major conclusions are insensitive to such differences. There are three major points to be made about these estimates. The first is that the stock of *R&D* capital contributes significantly to the explanation of cross-sectional differences in productivity and there is little evidence of a decline in its coefficient over time.[2] There is a minor rise in the estimated coefficient from 1967 to 1972 and a somewhat larger but not really significant decline from 1972 to 1977. Given this particular measure of *R&D* capital, based on a

[2] Here and subsequently, all statements about statistical "significance" should not be taken literally. Besides the usual issue of data mining clouding their interpretation, the "samples" analyzed come close to covering completely the relevant population. Tests of significance are used here as a metric for discussing the relative fit of different versions of the model. In each case, the actual magnitude of the estimated coefficients is of more interest than their precise "statistical significance."

TABLE 2—NSF-CENSUS STUDY: CROSS-SECTIONAL PRODUCTION FUNCTIONS, LOG VALUE-ADDED DEPENDENT VARIABLE[a] U.S. FIRMS: 1967; 1972; 1977

	(1)		(2)		(3)[b]	
Variables	1967	1972	1972	1972	1972	1977
ln Employment	.604	.622	.623	.586	.578	.611
	(.045)	(.046)	(.035)	(.038)	(.038)	(.039)
ln Capital Services	.224	.199	.161	.234	.254	.291
	(.041)	(.044)	(.032)	(.036)	(.036)	(.035)
ln *R&D* Stock (*db*)	.113	.135	.165	.126	.115	.089
	(.023)	(.026)	(.019)	(.019)	(.018)	(.017)
Basic Research (*BR*)	.396	.340	.274	.499	.517	.401
	(.240)	(.261)	(.215)	(.191)	(.189)	(.189)
Company-Financed Research (*FP*)	.190	.247	.068	.133	.138	.044
	(.097)	(.106)	(.100)	(.088)	(.088)	(.084)
N	386	386	652	491	491	491
SEE	.312	.336	.390	.312	.309	.290

Notes: ln Employment = log (total employment-employment of scientists and engineers); ln Capital Services = log of (depreciation plus interest on net assets plus machinery and equipment rentals); ln *R&D* Stock (*db*) = log of the "stock" of total *R&D* expenditures based on a 15 percent per year declining balance depreciation assumption; *BR* = basic research as a fraction of total *R&D*; 1972 in the 1977 equation, 1967 in 1967 and 1972. *FP* = fraction of *R&D* stock "private," company-financed *R&D* stock as a ratio to the total *R&D* stock, as of *t*. All equations include also a constant term and industry dummies. The number of industry dummies used depends on the data set and varies between 18 and 20. Standard errors are shown in parentheses.

[a]Value-added and materials used in research in 1967 and 1972.

[b]Value-added only.

15 percent per year declining balance depreciation formula (the results are insensitive to the particular formula used), the implied average (at the geometric mean of the sample) gross rate of return to *R&D* investment rises in a similar fashion from .51 in 1967 to .62 in 1972 (in col. 1) and falls from .39 in 1972 to .33 in 1977 (in col. 3). In either case the estimated rate of return is quite high and there does not appear to be any dramatic fall in it over time.

The second major finding is the significance and rather large size of the basic research coefficient. It seems to be the case that firms that spend a larger fraction of their *R&D* on basic research are more productive, have a higher level of output relative to their other measured inputs, including *R&D* capital, and that this effect has been relatively constant over time. If anything, it has risen rather than fallen over time. Using the formulation of equation (2) implies a very high premium on basic versus the rest, a $\hat{\delta}$ of between 2.5 to 4.5, a several hundred percent premium on basic research. Before I explore the implications of this result, I want to examine other dimensions of these data and see whether similar effects can be observed there too.

The last major result of interest is the significant positive coefficient on the privately vs. federally financed *R&D* mix variable. This variable is of most import for the older more established firms in subsample 4 (Table 1) but its sign is consistent throughout, indicating a positive premium on privately financed *R&D*, or equivalently a discount as far as federally financed expenditures are concerned. Here the implied premium is smaller, between 50 and 180 percent, but still quite large.

All the above results were based on cross-sectional level regressions that are subject to a variety of biases, the main one being the possibility that "rich" successful firms are both more productive and can afford to spend more of their own money on such luxuries as *R&D* and especially the basic variety. One

TABLE 3—GROWTH RATE OF PARTIAL PRODUCTIVITY, 1966–77

Variables	$N = 911$			$N = 652$ (*with industry dummies*)	
Constant	.019	.009	.012	–	–
BTRD 6677	.107		.117	.119	
	(.014)		(.017)	(.016)	
BCRD 6677		.095			.106
		(.014)			(.015)
BR 72	.056	.056	.059	.035	.034
	(.017)	(.017)	(.019)	(.018)	(.018)
FP 72	.011	.019	.017	.022	.030
	(.005)	(.005)	(.006)	(.007)	(.007)
SEE	.0383	.0384	.0337	.0305	.0307

Notes: Dependent variable: *BPT* 6677 = trend growth rate of deflated sales minus the trend growth of total employment multiplied by the share of payroll in total sales. *BTRD* = trend growth of deflated total *R&D* expenditures; *BCRD* = same for company-financed *R&D* expenditures; *BR* = basic research expenditures as of fraction of total research expenditures; *FP* = ratio of company-financed *R&D* stock to total; *SEE* = residual standard error. All equations contain also a term reflecting the variance of *R&D* and terms representing the growth of physical capital: age composition and depreciation as of 1972.

can reduce somewhat the possibility of this type of bias by focusing on firm-growth rates, the changes that occurred, rather than on their levels. To the extent that firms have idiosyncratic productivity coefficients that may be also correlated with their accumulated *R&D* levels, considering growth rates is equivalent to doing a "within" firms analysis, one that eliminates such fixed effects from the analysis. The next two tables present, therefore, the results of analyzing the growth in the partial productivity of these same firms during the whole 1966–77 period.

Table 3 presents the results of estimating partial productivity equations in the largest possible sample for which 1966–77 growth rates were computable ($N = 911$) and in the subsample with a successful 1972 Census match. Here again we find my three main results confirmed: the *R&D* growth term and the two mix variables, the basic research ratio, and the fraction of research financed privately all contribute significantly to the explanation of productivity growth.

On the assumption that the growth rate in the stock of *R&D* is roughly proportional to the growth in deflated *R&D* itself, the coefficient of *BTRD* should be estimating the same number as the coefficient of the *R&D* stock variable in Table 2. The results are in fact surprisingly close: about .12 in Table 3 as against .09 to .17 in Table 2. Moreover, there seems to have been no decline in this coefficient relative to the earlier 1957–65 period. In my previous study (1980a), I estimated the same coefficient to be .073. In the current replication and extension of this sample a similar equation for 1957–65 yields a *BTRD* coefficient of .086. Thus, if anything, the coefficient of *R&D* went up between the early 1960's and the early 1970's.

The second major finding of interest is the positive and significant basic research coefficient. It is hard to interpret its magnitude since the approximation outlined in equation (3) breaks down when the average growth rate of deflated *R&D* and of basic *R&D* is close to zero or negative. Consider, however, the following illustrative calculation. Raising the *BR* ratio by one standard deviation, from .026 to .097 at the mean, would increase the rate of growth of partial productivity by close to half a percent per year ($.071 \times .059 = .0042$). This same increase would raise the growth of total *R&D* by .107 for one year and would contribute a

TABLE 4—GROWTH RATE OF PARTIAL PRODUCTIVITY, BY INDUSTRY, 1966–77 (MATRIX 6, TOTAL $N = 991$)

Coefficients	Coefficients by the Estimated t-Ratio			
of	< -1.5	$-1.5-0$	$0-1.5$	$1.5+$
BTRD		2	7	10: Miscellaneous, Industrial Chemicals, Drugs, Stone & Glass, Machinery, Electronics, Electrical Equipment, Transportation Equipment, Scientific Instruments, Non-Manufacturing
*BR*72		5	8	6: Wood & Paper, Other Chemicals, Oil, Machinery, Aircraft, Non-Manufacturing
*FP*72	2	6	7	4: Oil, Rubber, Electronics, Aircraft

Notes: All equations contain also a term reflecting the variance of *R&D* and terms representing the growth of physical capital: age composition and depreciation as of 1972.

once-and-for-all increase in the level of productivity of .0125. Discounting the more "permanent" effect of basic research by a real interest rate of .05 yields an "equivalent" one-year effect of .084, or a 7 to 1 ratio in favor of basic research! If one allows for industry dummies which in this formulation represent separate industry trend rates of disembodied technical change, the effect of basic research is cut by about 50 percent, implying perhaps that a signficiant fraction of the estimated effect comes from spillovers that diffuse throughout the industry. Note that it is the only coefficient that is affected substantively when separate industry dummies are allowed for. Nevertheless, even a 3.2 to 1 ratio is quite high!

The third finding is the significant positive premium on company-financed *R&D*. Here too the implied premia are quite high, but given that the mix variable is defined in terms of stocks rather than flows, the calculations are more cumbersome. Consider starting from a zero growth position and a .7 ratio of private to total *R&D* stock. To move this fraction from .7 to .75, one would need to raise the private stock by 29 percent and the overall stock by 20 percent (without reducing absolutely the stock of federally financed *R&D* capital). There are different possible investment paths that would achieve this goal and would have somewhat different present value consequences. If one roughly doubled the rate of privately financed *R&D* expenditures, from the previous replacement level of .105 ($.7 \times .15$) to .205, one could achieve this target in slightly over two years. Ignoring discounting, this would lead to a once-and-for-all growth in productivity of .024, due to the growth in the total stock of *R&D and* a .0011 permanent increase in the rate of growth due to the shift of the fraction private ratio from .7 to .75. The present value of this second term is about .022, or of the same order of magnitude as the first term. That is, raising the stock of *R&D* by 20 percent but shifting it all into the private component doubles the effect of such dollars.

There are problems, however, with such an interpretation. If private *R&D* expenditures contribute more to productivity growth, one might have thought that when they are substituted for the total *R&D* growth measure, they might fit better and also have a higher coefficient. But that is not the case as can be seen from the results presented in columns 2 and 5 of Table 4. The total *R&D* measure does a little bit better both in terms of fit and in the overall size of its coefficient, implying that the contribution of federal dollars is not zero. That is perhaps what one should expect. Most of the direct output of federal research dollars is "sold" back to the government at "cost plus" and is unlikely to show up as an increase in the firm's own productivity. Thus all that one could expect to measure here are the within-firm spillover effects of such expenditures. What we may be detecting is that such effects are indeed present and positive, but we should not have

expected them to be of the same order of magnitude as would be the case for the firm's own investments in improving its productivity or profitability.

There are a number of econometric questions that can be raised about the robustness and sensitivity of such results. I will discuss only a few of these here. The most obvious question arises from the fact that even though I allowed, in the growth rates version, for separate firm intercepts and different industry trends, I am still assuming common *R&D* and the conventional capital coefficients across rather different industries. This is done from necessity rather than as a virtue. Estimating the same models industry by industry reduces the sample sizes drastically and raises greatly the relative noise level, making it rather hard to interpret the resulting estimates. Nevertheless, these estimates, which are summarized in Table 4, are quite consistent with the earlier story: 17 out of the separately estimated 19 coefficients for the *R&D* growth variable are positive and more than half of them are statistically significant at conventional significance levels. Similarly, the coefficients of the basic research ratio variable are positive in 14 of my 19 industries and signficant in over a third of them. The fraction private variable is less robust to the division of the sample into industries, with more than half of the coefficients still positive, but only 4 of them are statistically significant within particular industries. Two of these industries are indeed the ones where one would expect to find such an effect, aircraft and electronics, industries where the bulk of federal monies is spent. Nevertheless, it seems that the effect that is being caught by the fraction private variable has an important industry component, something that had been already noted in my earlier study (1980a), as does also the effect associated with the basic research variable, though to a lesser extent.

A number of other versions were computed using the growth in capital services rather than the depreciation and age composition variables that had been used to keep the results comparable to the earlier study, and the growth in *R&D* "capital" rather than the flow (and also different definitions of such capital). I also estimated versions using the "intensity" form for the *R&D* variable, to make it more comparable to other studies in the literature (my paper with Lichtenberg; Mansfield; and others).[3] By and large the results of these alternatives were somewhat weaker but not substantively different. Perhaps the most interesting alternative estimate is the intensity version using the growth of capital between 1967 and 1972 as its capital measure:

$$(4)\quad BPT6677 = \ldots \underset{(.069)}{.243ACRS} + \underset{(.024)}{.045ABR} + \underset{(.130)}{.180DLCS} \qquad SEE = .0316 \quad \text{(Subsample 4)}$$

where *ACRS* is the average company *R&D* to sales ratio, averaged over 1967 and 1972, *ABR* is a similar average basic to total *R&D* ratio, and *DLCS* is the rate of growth in deflated capital services between 1967 and 1972. This version is closest in form to the equation estimated by Mansfield on much smaller samples. The basic results are similar, however. Basic *R&D* is a signficant contributor to productivity growth with an implied basic to company premium of about 5 to 1 (given an average *R&D* to sales ratio of .035).

The final set of results to be presented here, in Table 5, relate to the relative profitability of our firms in 1972 and 1977. The dependent variable, *GRR*, is the ratio of gross profits (value-added minus labor costs and plus *R&D*) to total gross fixed assets. The independent variables include the ratio of *R&D* capital (undepreciated) to total fixed assets and our ubiquitous *R&D* mix variables: the basic research and fraction private

[3]The intensity version uses the fact that $\alpha = (\partial Q/\partial K)K/Q$ and reexpresses $\alpha\dot{K}/K$ as $\rho[R/Q]$, where $\rho = \partial Q/\partial K$ is the marginal product (gross rate of return) of *R&D* capital and it has been assumed that $\dot{K} = R - \delta K \simeq R$, i.e., either $\delta \sim 0$ (no depreciation) and/or initial K very small. This formulation has the advantage that it does not impose the assumption of a constant *elasticity* across different firms, replacing it instead by the, possibly more plausible, assumption of the constancy of rate of return.

TABLE 5—GROSS PROFIT RATE REGRESSIONS
(GRR = (Value-Added-Payrolls + R&D)/Gross Assets))

Dependent Variable and Sample Size	Constant	Coefficients of: R&D Capital to Total Fixed Assets Ratio	Basic R&D Ratio	Fraction Private	SEE
GRR 72					
$N = 652$ (a)	.144	.088	.344	.107	.262
	(.049)	(.012)	(.144)	(.048)	
(b)		.060	.187	−.012	.237
		(.013)	(.138)	(.052)	
$N = 491$ (a)	.117	.080	.514	.154	.264
	(.052)	(.013)	(.139)	(.051)	
(b)		.061	.366	.074	.227
		(.015)	(.138)	(.057)	
GRR 77					
$N = 491$ (a)	.341	.031	.402	.033	.313
	(.064)	(.019)	(.187)	(.068)	
(b)		.004	.261	−.028	.292
		(.022)	(.187)	(.077)	

Notes: (a) Regressions do not contain industry dummies; (b) do.

ratios. Even though the dependent variable is quite different, the overall results are rather similar to the earlier ones. The *R&D* capital variable is positive and almost always statistically significant through its coefficient is a bit low if it is to be interpreted as a rate of return to it. The basic research variable is both large and significant though possibly too large to be credible. Given that the ratio of total *R&D* capital to total fixed assets is only about .05 on average, the 1972 coefficients imply a δ of about 30 to 60. The fraction private ratio also contributes positively to profitability but its effect largely disappears once industry differences are allowed for. The results for 1977 are weaker than those for 1972, the residual variance is signficantly higher, but they too suggest the importance of basic research even in this context.

A similar analysis was performed using an estimate of the net rate of return as the dependent variable, subtracting depreciation from the numerator of *GRR* and using a net stock concept for the denominator and also in the definition of the *R&D* capital variable. While the fit was significantly worse when using this definition of the dependent variable, the overall results were rather similar. The net return version was also available for 1967 and the results using it indicate a relatively constant and significant coefficient for the basic research ratio while the coefficient of the total *R&D* stock rises from 1967 to 1972 and than falls again in 1977 (from .11 to .16 and down to .06). It is doubtful whether these fluctuations represent real trends or, more likely, reflect the larger noise level in the 1977 data and the changing composition of these samples. In any case, the profitability regressions are consistent with the productivity level and productivity growth rate based results described earlier (Tables 2, 3, and 4).

III. Discussion and Summary

There are three major findings in this paper: *R&D* contributed positively to productivity growth and seems to have earned a relatively high rate of return; basic research appears to be more important as a productivity determinant than other types of *R&D*; and privately financed *R&D* expenditures are more effective, at the firm level, than federally financed ones. These findings are not entirely new. The first finding has been documented in a number of earlier studies

(see my 1980a, b papers; my article with Mairesse, 1984; A. N. Link, 1981a; and others). What is new in this paper in this regard is a confirmation of this finding on a much larger and more recent data set. It also presents evidence for the view that this effect has not declined significantly in recent years, in spite of the overall slowdown in productivity growth and the general worry about a possible exhaustion of technological opportunities.[4]

The evidence for a "premium" on basic research is much more scarce. The major previous paper suggesting this type of a result is Mansfield which uses aggregate data for 20 industries for 1948–66 and data for 16 firms during 1960–76, and finds a signficant premium on basic research, on the order of 2 to 1 at the industry level and 16 to 1 at the firm level. (See also Link, 1981b, for similar results for 1973–78 based on data for 55 firms.) In this paper I get similar though somewhat smaller effects at the firm level, using a much larger and more representative sample. I also find that differences in levels of productivity and profitability are related to differences in the basic research intensity of firms.

Such findings are always subject to a variety of econometric and substantive reservations. In this context the two major related issues are simultaneity and the question of how major divergences in private rates of return persist for such long periods. It is possible to argue that it is not *R&D*, or its basic research component, that causes firm "success" as measured by productivity and profitability, but rather that success allows firms to indulge in these types of luxury pursuits. It is difficult to argue about causality on the basis of what are essentially correlational data. It is possible to use simultaneous equation techniques to estimate such models, but then the argument shifts to the validity of the exogeneity assumption for the particular instruments. In the context of my specific data set, it is hard to think of any valid instruments except for possibly lagged values of the same variables, which raises some problems of its own. The best evidence for the notion that these results are not entirely spurious is provided by the growth rates where the individual firm levels are partialed out of the analysis. But, here too, one could argue about the impact of common unanticipated "luck" elements. Unfortunately, it is unlikely that one could use lagged growth rates as instruments, since there is very little correlation in growth rates over time at the firm level. While an attempt will be made in further work with these data to estimate more extended simultaneous equations versions of such models, I am not too optimistic as to what can be accomplished in this regard. The evidence presented here should not be interpreted as "proving" that *R&D*, and especially its basic component, are important for productivity growth but rather as presenting some prima facie evidence in support of such an interpretation. In this sense it is an exercise in economic rhetoric (Donald McCloskey, 1983).

It is even more difficult to respond to the theoretical a priori argument that such results cannot be true since they imply widely differing rates of return to different activities under the control of the same firm. One's response to this depends on one's views as to the prevalence of equilibria in the economy. While it is likely that major divergences in rates of return are eliminated or reduced in the long run, the relevant runs can be quite long. *R&D* as a major component of firm activity was undergoing a diffusion process in the 1950's and 1960's and may not have reached full equilibrium even by the end of our period. This may be especially true of the basic research component where the risks are much greater and the uncertainty introduced

[4]The finding that the coefficients in a logarithmic regression have not declined over time does not dispose of the possibility that there could have been an overall loss in accumulated knowledge capital due to accelerated obsolescence. A proportional decline in the effectiveness of past capital or in the rate that *R&D* is converted into new knowledge capital need not show up as decline in the slope coefficient, it would get absorbed into the shifting constant. Disproportionate shifts should, however, have an impact on the estimated slope coefficient. Also, a pure obsolescence shock to old knowledge capital would have called forth an increase in the rate of *R&D* expenditures, something which has not been observed in the data. I am indebted to M.N. Baily for this point.

by changing government policies and the changing economic environment make it quite difficult to decide what is the right level for it.

A somewhat different version of this argument would claim that the world is indeed in approximate equilibrium but that different firms face different opportunities for doing research, basic or otherwise, are in different ecological niches, and hence have different coefficients in their "production functions." This would explain why different firms are observed to spend different amounts on *R&D* while actually earning about the same rate of return on it. When a constant coefficients production function is fit to such data, it will fit because it is approximating a market equilibrium relation. If the level of *R&D* invested were independent of the coefficient, then such a function would just reproduce its average share and not produce any evidence of excess returns. But if, as is reasonable, *R&D* is invested optimally with firms which have better opportunities, higher coefficients, investing more, this will induce a positive correlation between *R&D* and its individual coefficient and lead to an upward bias in the estimated "average" coefficient.[5] The resulting "larger" coefficient, larger than the observed factor share, will be interpreted, wrongly, as implying a higher rate of return than is actually prevailing at the individual level.

This argument may be recognized as a version of the earlier attacks on the Cobb-Douglas production function combined with a random coefficients interpretation of the same phenomenon. In its extreme form it is testable. Since there are time-series data available for individual firms, one could try to estimate individual firm parameters and check whether they are in fact distributed as is predicted by this particular argument. While individual parameters are unlikely to be well estimated, given the relative shortness of the available time-series, the parameters of the distribution of such coefficients might be estimable with more precision. I intend to pursue this possibility in future work.

To restate again the major points of the paper: a newly available body of data on all the major firms performing *R&D* in the United States has been examined and evidence has been presented for the proposition that *R&D* contributes significantly to productivity growth, that the basic research component of it does so even more strongly, and that privately financed *R&D* expenditures have a significantly larger effect on private productivity and profitability than federally financed *R&D*. These findings are open to a number of reservations. Nevertheless, they do raise the issue that the overall slowdown in the growth of *R&D* and the absolute decline in basic research in industry which occurred in the 1970's may turn out to have been very costly to the economy in terms of foregone growth opportunities.

[5]A positive correlation is not enough, by itself, for a positive bias. The weight of an individual firm slope coefficient in the cross-sectional estimate is proportional to the *square* of the deviation of *R&D* stock from its mean. A positive correlation between levels does not translate itself directly into a positive correlation between the level of one variable and the square of the other, except for certain skewed distributions. Since we do not observe the individual coefficients directly, it is rather difficult to check out this conjecture.

REFERENCES

Bound, John et al., "Who Does *R&D* and Who Patents?," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 21–54.

Griliches, Zvi, "Issues in Assessing the Contribution of *R&D* to Productivity Growth," *Bell Journal of Economics*, Spring 1979, *10*, 92–116.

______, (1980a) "Returns to Research and Development Expenditures in the Private Sector," in J. W. Kendrick and B. Vaccara, eds., *New Developments in Productivity Measurement*, NBER *Studies in Income and Wealth* No. 44, Chicago: University of Chicago Press, 1980, 419–54.

______, (1980b) "*R&D* and The Productivity Slowdown," *American Economic Review, Proceedings*, May 1980, *70*, 343–48.

______, "Comment" on Edwin Mansfield, "R and D and Innovation: Some Em-

pirical Findings," in Z. Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, 1984.

______ and **Hall, B. H.**, "Census-NSF *R&D* Data Match Project: A Progress Report," in *Development and Use of Longitudinal Establishment Data*, Economic Research Report, ER-4, Bureau of the Census, 1982, 51–68.

______ and **Lichtenberg, F.**, "*R&D* and Productivity Growth at the Industry Level: Is There Still a Relationship?," in Z. Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, 1984, 465–96.

______ and **Mairesse, J.**, "Comparing Productivity Growth: An Exploration of French and U.S. Industrial and Firm Data," *European Economic Review*, April 1983, *21*, 89–119.

______ and ______ ,"Productivity and *R&D* at the Firm Level," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 339–74.

Hall, Bronwyn H., "Historical *R&D* Panel: 1957–77. Public Use Correlation, Matrices Tape Documentation," unpublished, September 1984.

Link, A. N., (1981a) *Research and Development Activity in U.S. Manufacturing*, New York: Proeger, 1981.

______, (1981b) "Basic Research and Productivity Increase in Manufacturing: Additional Evidence," *American Economic Review*, December 1981, *71*, 1111–12.

McCloskey, Donald N., "The Rhetoric of Economics," *Journal of Economic Literature*, June 1983, *22*, 481–517.

Mansfield, Edwin, "Basic Research and Productivity Increase in Manufacturing," *American Economic Review*, December 1980, *70*, 863–73.

Pakes, Ariel and Schankerman, M., "The Rate of Obsolescence of Knowledge, Research Gestation Lags, and the Private Rate of Return to Research Resources," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 209–32.

National Science Foundation (Jaffe, S.A.), "A Price Index for Deflation of Academic *R&D* Expenditures," NSF 72–130, Washington: USGPO, 1972.

______, *Trends to 1982 in Industrial Support of Basic Research*, NSF 83–302, Washington: USGPO, 1983.

______, *National Patterns of Science and Technology Resources*, NSF 84–311, Washington: USGPO, 1984.

Golden Parachutes, Shark Repellents, and Hostile Tender Offers

By CHARLES R. KNOEBER*

The tender offer has recently become a popular way to change control over the assets of a firm.[1] Unlike the more traditional merger proposal which must be approved by the target firm's board of directors before it is submitted to a shareholder vote, a tender offer is made directly to the target's shareholders and requires neither approval nor even notification of the target's board of directors. While tender offers and mergers each evoke controversy, a simple argument suggests either is beneficial to both the acquiring firm and to the shareholders of the acquired (target) firm. Each is a voluntary exchange and so would be agreed to only if both parties to the exchange expect to benefit. Since mergers require approval of the management of the target firm, however, they add another party to the exchange.[2] Again invoking the argument of mutually beneficial exchange, it must be that, for mergers, not only do the acquiring firm and shareholders of the target benefit but so do the managers of the target. Since target management is not a party to the exchange embodied in a tender offer, though, it need not benefit and indeed may be harmed. This explains why tender offers are sometimes hostile (opposed by management of the target) and sometimes friendly, while mergers are always friendly.

The noninvolvement of target management in a tender offer has further been argued to be a desirable feature of this form of transferring control over corporate assets, since it may provide the additional benefit of displacing poorly performing management. Indeed, the threat of a tender offer is seen as inciting managers to better performance and enriching share-holders (Henry Manne, 1965; Frank Easterbrook and Daniel Fischel, 1981). This argument suggests tender offers are more desirable than mergers, and that recent tender offer experience is an improvement over the previous almost complete reliance on mergers to alter control over corporate assets.

Despite this, much recent activity can be viewed as attempting to discourage hostile tender offers. The Williams Act (1974) provides federal rules regulating tender offers. These rules expand the period during which a tender offer must remain open, dictate substantial disclosure as to the bidders' plans and sources of finance, and contain antifraud provisions that provide target management with the standing to sue for injunctive relief (see Gregg Jarrell and Michael Bradley, 1980). The effect is to better enable target management to obstruct hostile tender offers and so reduce the advantage such offers have over the more traditional merger. Of more interest here are two voluntary measures adopted by a firm's shareholders. The first of these has been labeled "golden parachutes."

*Associate Professor of Economics and Business, North Carolina State University, Raleigh, NC 27695. Many people have assisted me with useful criticisms. I especially thank Steve Allen, David Ball, Clive Bull, Dwight Grant, Mark Fisher, Jim Hess, Jack Hirshleifer, Steve Margolis, Wally Thurman, and an anonymous referee.

[1] In 1960, only 7 tender offers were made for shares in U.S. firms. This represented about four-tenths of 1 percent of all announcements of mergers or acquisitions involving U.S. firms. By 1970, the number of tender offers was 34 or about 2 percent of all acquisition announcements (Patrick Davey, 1977). In 1975, tender offers numbered 71 or 7 percent of acquisition announcements. An apparent high was reached in 1977 with 181 tender offers representing 15 percent of acquisition announcements. The figures for 1980 and 1982 were 123 and 94 tender offers which were 8 and 4 percent of acquisition announcements for the respective years (Douglas Austin, 1980; Austin and Michael Jackson, 1984).

[2] It is the board of directors that must approve a merger proposal before it is submitted to shareholders. While managers are represented on the board (inside directors), they need not control board decisions. So merger proposals opposed by management may be submitted to shareholders.

These are contractual agreements with management that provide substantial (often millions of dollars) payments to managers who elect to (or are forced to) leave the firm when a change of control takes place.[3] A survey of 665 industrial companies in 1982 found 15 percent provided golden parachutes to top management (Ann Morrison, 1982). The obvious effect of this measure is to insulate a target's management from harm even in the case of hostile tender offers. Thus, the advantage of a tender offer over a merger (its policing effect on management) is lost.

The second voluntary measure is the adoption (by shareholder vote) of amendments to corporate charters or bylaws which discourage tender offers. These so-called "shark repellent" amendments provide for super majority voting on mergers and sale of assets, stagger the terms of corporate directors, and impose other impediments to hostile tender offers (for descriptions, see Ronald Gilson, 1982; Harry DeAngelo and Edward Rice, 1983; Scott Linn and John McConnell, 1983). As in the other instances, shark repellents act to dilute the advantage of tender offers over traditional mergers. In all cases, the effect is to make hostile tenders less likely, or to reduce the possibility of an exchange of control of corporate assets not agreed to by current management.

Those who believe in the beneficial effect of hostile tender offers on manager performance typically deplore these recent actions which discourage hostile offers. Most particularly, critics have attacked golden parachutes and shark repellents. The criticisms have led to several proposals to regulate such actions. A legislative rule amending the Williams Act, proposed by Lewis Lowenstein (1983), aims to restrict the actions management can take when confronted with a tender offer. A somewhat similar judicial rule requiring management passivity in the face of tender offers has been proposed by Easterbrook and Fischel. Most recently, an advisory committee to the SEC has recommended rules prohibiting shark repellent amendments to corporate charters and bylaws and restrictions on the use of golden parachutes (*Wall Street Journal*, March 1984).

The fundamental paradox in this attack on obstructions to hostile tender offers is that it is based upon a rejection of the very argument that is used to defend the desirability of any tender offers. This argument is the mutual benefit of voluntary exchange. Tender offers are desirable because they are voluntary and so both the acquiring firm and shareholders of the target benefit. Golden parachutes and shark repellents are viewed as undesirable because they work to the detriment of shareholders. They are, however, voluntarily agreed to by shareholders. How, then, can agreement by these shareholders not be taken as evidence of shareholder benefit (when it is in the case of tender offers)?

The object of this paper is to examine the contractual relation between shareholders and managers, how tender offers (from an outside party) affect this relation, and to suggest that it may well be in shareholders' as well as managers' interest to agree to restrict the possibility for outsiders to disrupt their relation with a hostile tender offer. Quite simply, the same argument used to advocate tender offers can also be made to advocate voluntarily adopted restrictions on hostile tender offers. Doing so casts a considerably different light on recent proposals to regulate such actions as the use of golden parachutes and shark repellent amendments to corporate charters.

The organization of the paper is as follows. Section I characterizes the relation between manager and shareholders and the nature of the contracts which might be expected between the two. Section II considers the effect of hostile tender offers on this contractual relationship and suggests a possible beneficial role for golden parachutes and shark repellents. Section III focuses empirically on golden parachutes. Here, the earlier sections are used to construct hypotheses about which firms will provide golden parachutes to managers and about

[3]Shareholders do not vote explicitly to provide managers with golden parachutes. They are, however, notified of such contracts in proxy statements and so can be viewed as approving them by acquiescence.

the relation between tenure and compensation for managers of firms providing golden parachutes. These hypotheses are then tested on a sample of 331 firms.

I

Shareholders of a firm employ a manager as their agent to make decisions regarding the use of the firm's resources. Shareholders are presumed to be risk neutral and interested only in maximizing the value of their shares (their wealth). The manager is a utility maximizer interested both in his money income and his on-the-job consumption. Denote this consumption as a, where a is meant to include activity typically designated as shirking as well as consumption of job perquisites. Consumption a by the manager imposes a cost on shareholders. Consequently, define units of a in terms of their cost to shareholders; each unit entailing a \$1 reduction in the (combined) wealth of shareholders. Some on-the-job consumption is desirable in that the value to the manager exceeds the cost to shareholders. That is, shareholders will want the manager to consume on the job up to the point (call this a^*) where the marginal benefit, MB, of a to the manager just equals the marginal cost (\$1) to shareholders. The reason is that compensation in kind, a, is an alternative to compensation in dollars and shareholders minimize their cost of hired management if they choose the cheapest form of compensation.[4] Until $a = a^*$, on-the-job consumption is a cheaper form of compensation than dollar payments.

The manager, however, chooses a and so may select on-the-job consumption greater than that which shareholders desire. If so, some on-the-job consumption costs shareholders more than it is worth and the difference I will call agency cost or the cost of improper incentives provided the manager.[5] That is, agency cost, A, is defined as

$$A = \int_{a^*}^{a} (1 - MB(a))\, da. \tag{1}$$

In order to reduce these agency costs or provide correct incentives to the manager, shareholders must somehow tie the manager's dollar compensation to his on-the-job consumption. They must impose a price for on-the-job consumption. (The market for corporate control might also perform this function. The implicit assumption here is that the only discipline imposed on a manager is that of the shareholders. In the empirical section of the paper, this assumption is relaxed.) If a can be measured exactly by shareholders and so is known by both managers and shareholders, then the price can be set equal to \$1 and agency costs are eliminated. For example, the wage contract with the manager might take the following linear form

$$W = \psi_1 - a, \tag{2}$$

where ψ_1 is the fixed salary component and a \$1 penalty is imposed for each unit of a chosen by the manager. If a cannot be measured without error by shareholders, a contract such as (2) may still eliminate agency costs; but only in a special case. Say the shareholders' estimator of a, designated $\hat{a}$, is unbiased,

$$\hat{a} \sim a, \quad \sigma_{\hat{a}}^2. \tag{3}$$

If the manager is risk neutral, then a contract like (2) with the shareholder estimator of on-the-job consumption replacing actual on-the-job consumption will again lead the manager to choose $a = a^*$ and eliminate agency costs.

The actual case of managers and shareholders, however, entails both an inability of shareholders to measure exactly on-the-job consumption by the manager and manager risk aversion. Here, the primary problem in principal-agent relationships must be faced. Better incentives provided to the manager entail greater risk bearing by the manager. For example, if a contract like (2) is slightly

[4] For a more complete discussion, see Harold Demsetz (1983).

[5] Agency cost here includes only a portion of what Michael Jensen and William Meckling (1976) refer to as agency cost.

generalized,

$$W = \psi_1 - \psi_2 \hat{a}. \tag{4}$$

As ψ_2, the price imposed for units of on-the-job consumption, is increased to provide better incentives, the variance of the manager's wage is also increased. Specifically,

$$\sigma_W^2 = \psi_2^2 \sigma_{\hat{a}}^2, \tag{5}$$

so

$$d\sigma_W^2 / d\psi_2 = 2\psi_2 \sigma_{\hat{a}}^2. \tag{6}$$

Now there is a cost of providing better incentives (reducing agency cost) in that additional manager risk bearing is required and the manager must be compensated for such risk bearing to induce him (or her) to accept the contract. Given some $\hat{a}$, the problem faced by the shareholders is to find a contract that maximizes their own wealth, subject to the constraints that the manager chooses a to maximize his expected utility when he faces the contract, and that the manager must receive sufficient compensation that his expected utility is no less than that which he could receive in some other employment. This optimal contract will be such that the marginal risk cost from altering the price (not necessarily a linear price) of measured on-the-job consumption is just offset by the induced change in agency cost.

Define SW as shareholder wealth when this optimal contract is selected. Holding constant manager preference for on-the-job consumption and manager attitude toward risk, SW is a function of the precision of the shareholder estimator of manager performance (on-the-job consumption), $SW(\sigma_{\hat{a}}^2)$, where

$$dSW / d\sigma_{\hat{a}}^2 < 0. \tag{7}$$

This is due to the fact that as $\sigma_{\hat{a}}^2$ becomes larger, shareholders must either incur greater agency cost or a greater wage payment or both to compensate the manager for additional risk bearing.

Now assume there are two possible estimators of manager performance. The first, $\hat{a}_1$, employs only current information (say the current performance of the firm) to estimate manager performance. The second, $\hat{a}_2$, employs current information and future information to estimate manager performance. The first estimator cannot be more precise than the second and will likely be (perhaps much) less precise.

$$\sigma_{\hat{a}_1}^2 \geq \sigma_{\hat{a}_2}^2 \tag{8}$$

Consequently,

$$SW(\hat{a}_1) \leq SW(\hat{a}_2). \tag{9}$$

Shareholders, then, would generally prefer to use the second estimator of manager performance.

One way to do this would be to initially reward the manager using $\hat{a}_1$ as an estimator of his performance, and then to settle up in the future by making additions or subtractions from this initial payment as better future information becomes available. The difficulty with this method is that the manager must remain in the employ of the shareholders to receive the settling-up increments and decrements. If managers are free to quit at their discretion, as indeed they are, then a manager knowing that shareholders underestimated a initially (recall the manager chooses a and so knows the extent of mismeasurement) will have an incentive to quit.[6] The manager takes the money and runs. The possibility of such opportunistic behavior by the manager will discourage shareholders from agreeing to such a settling-up contract initially.

Another contract that allows the use of $\hat{a}_2$, but is not susceptible to opportunism by the manager, entails a small payment initially based on the estimator $\hat{a}_1$ and then a nonnegative settling-up bonus paid in the future when better information becomes available.[7]

[6] To the extent other employers discover that opportunism occurred (learn the true a), the market for managers may penalize and so discourage such behavior by managers. See Eugene Fama (1980).

[7] To ensure that the settling-up bonus is nonnegative, the initial payment may need to be negative. If so, the wealth of the manager becomes a constraint on the use of this sort of contract.

This contract that involves deferred compensation removes the incentive for the manager to behave opportunistically, since the deferred compensation acts as a bond (or precommittment) tying the manager to continued employment.[8]

These arguments suggest that in situations where a manager is risk averse and shareholders cannot determine manager performance exactly but can improve their estimator of manager performance as time passes, an optimal contract with the manager will include deferring some expected compensation to be paid in the future.[9] The greater the value of future information or the larger the difference between $\sigma^2_{\hat{a}_1}$ and $\sigma^2_{\hat{a}_2}$, the greater will be the share of compensation which will be deferred.

Beyond this, the actual amount of deferred compensation is not known, since it depends on information that only becomes available in the future. A contract cannot be written explicitly stating the amount of deferred compensation. Indeed, it is unlikely that a contract can even be written that specifies how this compensation will be determined in the future. This is due to the variety and complexity of information which may become available. Detailing all the future possibilities and contingent payments will at the least be expensive and very likely futile. Consequently, such a long-term deferred compensation contract will be largely implicit.[10]

Obviously, such a contract presents an avenue for shareholders to behave opportunistically. They may fail to pay the deferred compensation and perhaps even discharge the manager. Shareholders may take the money and run. This is not likely, however. The decision to pay the deferred compensation rests with the board of directors. Should they renege on the implicit contract (and should this be known by other managers), the firm would find it difficult to retain or replace managers. The cost of behaving opportunistically would outweigh the gains.[11] The implicit contract would be self-enforcing (see Lester Telser, 1980; Benjamin Klein and Keith Leffler, 1981) and the manager could trust that he will indeed be paid deferred compensation as due.

Optimal contracts between a manager and shareholders (via the board of directors) will often involve deferring compensation until better information about manager performance becomes available. These contracts will necessarily be long term and likely be implicit.[12] However, opportunistic behavior will not be a problem. The manager's deferred compensation acts as a bond, deterring opportunism on his part, and the cost of developing a reputation for unreliability (among other managers) deters opportunism on the part of the board of directors.

II

What effects will tender offers, particularly hostile ones, have in this contracting framework? Importantly, they provide another

[8] Related arguments as to the advantage of deferred compensation and long-term contracts are made by Gary Becker and George Stigler (1974), Jonathan Eaton and Harvey Rosen (1983), and Richard Lambert (1983).

[9] The primary reason for deferring compensation here is to allow a more precise determination of manager performance. This reason is not necessary for compensation to be deferred and has not received primary attention in other studies. For example, Edward Lazear (1979, p. 1272) while acknowledging this monitoring rationale for deferring compensation, develops a model that generates deferred compensation under the assumption that performance is observed perfectly and immediately. This perfect measure of performance allows a worker to contract for an optimal level of performance (a^* in my approach), and deferred compensation arises as a mechanism to ensure the worker abides by the contract.

[10] Even if the contract can be written, if courts (third-party enforcers) cannot observe the contingencies stated in the contract, as is likely, then an explicit contract will be unenforceable and so equivalent to an implicit contract. See Clive Bull (1983).

[11] I presume that firms never find it desirable to renege because of these reputation costs. The present value of the cost of behaving opportunistically, however, is lower when a firm's discount rate is higher. Consequently, firms become more likely to behave opportunistically as this discount rate rises. See Bengt Holmstrom (1983) and H. Lorne Carmichael (1984).

[12] As an example, General Motors recently provided bonuses exceeding $1 million to each of its five highest ranking managers. These bonuses are deferred compensation (the previous three years "GM's compensation fell 'way below' other companies": *Wall Street Journal* April 16, 1984, p. 8), paid on long-term, implicit contracts. Their size and timing were not dictated by explicit contracts.

avenue for shareholder opportunism. A tender offer bypasses the board of directors and appeals directly to shareholders to sell their shares. Once control has changed hands, a manager may be discharged or, if retained, not paid deferred compensation due. The acquiring firm then appropriates this delayed compensation (the prospect of which may be partially responsible for a premium over current stock price paid to tendering shareholders of the acquired firm). Reputation does not prevent shareholders from participating in such opportunism (tendering their shares), as they are essentially anonymous. Nor does reputation deter the bidding firm from such behavior. While this firm would not treat its own managers in such a fashion, developing a reputation for opportunistic behavior in takeovers need not destroy a reputation for reliability in dealing with the firm's own managers ("honor among thieves").

Importantly, the possibility of opportunistic behavior that arises with tender offers is absent with merger proposals. Such proposals must be approved by the board of directors of the acquired firm who have an incentive (to maintain individual reputations and so future employment prospects on other boards for outside directors;[13] obvious financial incentives for inside directors) to assure that any proposal forwarded to shareholders preserves the deferred compensation due management.

This suggests a new explanation for golden parachutes and shark repellents. Both can be viewed as attempts to eliminate the possibility of opportunism toward managers with implicit long-term deferred compensation contracts. A golden parachute is simply a bond posted by shareholders which accrues to the manager should opportunism accompany a takeover. If the bond is sufficiently large, there is no incentive for opportunism to occur. The acquiring firm might be able to capture deferred compensation due the manager of the acquired firm but only by forfeiting the bond (golden parachute payment).[14] Similarly, shark repellent amendments make hostile tender offers more costly and less likely to succeed. The effect is to make it more advantageous to pursue a friendly tender offer (one approved by management). To gain such approval, the bidder must assure the current managers of the target that it will not behave opportunistically. Both golden parachutes and shark repellents, then, deter opportunistic behavior by bidding (raider) firms.

The advantage to current shareholders of a firm providing golden parachutes or adopting shark repellents is that by so doing, these shareholders can assure managers that implicit deferred compensation contracts will not be reneged. Without this assurance, managers would not agree to such contracts. They would require immediate compensation that would necessitate the use of a less precise measure of manager performance and so (see equation (9)) less shareholder wealth. These obstructions to hostile takeovers, then, allow better contracting between manager and shareholders. Voluntary trade is mutually beneficial.

The prospect of hostile tender offers may indeed impose discipline on a firm's managers as argued by Manne and so provide a benefit to shareholders. If so, there is a cost of restricting the possibility of hostile offers with golden parachutes and shark repellents. The arguments presented here, though, suggest there may also be an offsetting gain. These impediments to hostile tender offers enlarge the contracting possibilities available between managers and shareholders, and, by allowing more effective contracting, enrich shareholders. This casts golden parachutes and shark repellents in a new light. They may well be in the interests of shareholders. The case against these obstructions to hostile tender offers is no longer clear.

[13]See Armen Alchian (1984).

[14]If the golden parachute is "too large," exceeding the likely value of deferred compensation due, a manager may voluntarily leave the firm when control changes hands even if no opportunism occurs on the part of the acquiring firm. This suggests a limit to the size of the compensation provided by a golden parachute.

Indeed, since these devices are voluntarily adopted by shareholders, it would seem that, where adopted, the gains from restricting hostile tenders outweigh the costs. That is, voluntarily adopted restrictions (like any voluntary trade) are mutually beneficial. Some evidence already seems to confirm this. Linn and McConnell, using the techniques of empirical finance, found a weak positive effect on stock price (shareholder wealth) due to the adoption of shark repellent amendments to charters or bylaws.[15] A similar study by Richard Lambert and David Larker (1985) also found a positive effect on stock price due to the provision of golden parachutes. Further confirmation is provided in the following section where two sorts of tests are performed. The first examines the incidence across firms of golden parachutes. The second examines the relation between tenure and compensation, or the importance of implicit deferred compensation, for managers of firms which do and firms which do not provide golden parachutes.

III

The above arguments make two important points. First, shareholders have more to gain from the use of implicit deferred compensation contracts the poorer is current information about manager performance relative to that which becomes available later. Second, such contracts become less acceptable to managers the more likely is a hostile tender offer and the attendant possibility of opportunism. These points suggest which firms will be most likely to provide managers with golden parachutes—those firms for which poor current information about manager performance make implicit deferred compensation contracts desirable but that face management resistence due to fear of tender related opportunism. To explain empirically firms' decisions to provide golden parachutes, variables measuring the gain to waiting for future information to evaluate managers and variables measuring the likelihood of a hostile tender offer are required.

I begin with the second task. Empirical researchers have recently examined financial and product market characteristics of firms to discover which affect the likelihood of being acquired by either friendly or hostile means (Robert Harris et al. 1982; Steven Schwartz, 1982). A negative effect of size (as measured by assets) was found. Additionally, measures of the market value of outstanding stock relative to earnings (Harris et al.) or relative to book value of assets (Schwartz) were found to be negatively related to the likelihood of acquisition. Beyond this, no consistent effect of other firm or product market characteristics was found. These findings and the presumption that firms more likely to be acquisition targets are similarly more likely to face a hostile tender offer suggest the following hypotheses. Smaller firms (fewer assets) and firms with lower price-earnings ratios (following Harris et al.) will be more likely to provide golden parachutes to managers.

These hypotheses, though, depend upon the implicit assumption maintained in the first two sections of the paper, that shareholder-manager contracting is the only mechanism which disciplines managers. That is, they consider only the benefit of golden parachutes. Where both such contracting and the market for corporate control work to provide incentives to managers, there is a cost of providing golden parachutes because they impede the incentive effects from the market for control. Where tender offers are more likely, this market will be more effective at motivating managers and so the cost of golden parachutes will be higher. Where tender offers are likely to occur only if manager behavior is egregious, the market for control will be less effective at motivating managers and the cost of golden parachutes will be lower. Increases in firm size and price-earnings ratio reduce the likelihood of a tender offer. Consequently, the cost of golden parachutes should fall with firm size and price-earnings ratio, and they should become more common. This runs counter to the previously predicted negative relation between the existence of golden parachutes and

[15] A similar study by DeAngelo and Rice, however, found essentially no effect on stock price.

firm size and price-earnings ratio. The net effect, then, is unclear and may not be monotonic. For example, if the cost effect dominates for small firms, the incidence of golden parachutes could first rise with increased firm size (as the cost of golden parachutes falls), but then may decline again for very large firms. For these large firms, the cost of golden parachutes is small, but, since tender offers are very unlikely, managers need little assurance of shareholder reliability and so the gains from golden parachutes may be even smaller. The net effect of firm size and price-earnings ratio on the likelihood of a golden parachute, then, is of ambiguous sign and may not be monotonic.

An additional variable that may affect the likelihood of a successful hostile tender offer is the fraction of shares held by the manager (or perhaps the board of directors). The larger this fraction, the greater the proportion of other shares that must be tendered for a hostile offer to succeed, and so the less likely such an offer will succeed. As a consequence, golden parachutes are less useful. The resulting hypothesis is that golden parachutes are less likely, the greater the fraction of shares held by the manager (board of directors).

Returning to the first task, variables measuring the gain to waiting for future information must be constructed. Two general characteristics of firms seem to matter. The first is the noise (exogenous influences) encountered when measures of firm performance (outcome realizations) are used to impute manager performance. Waiting for more information allows some of the initial noise to be explained and so allows a more precise estimate of manager performance. The second is the existence of lags between manager performance and outcome realizations for the firm. Where important decisions by managers do not lead to measurable outcomes for several years, current measures of firm performance have little to do with current manager performance. Only by waiting for future information can firm performance be used to impute manager performance. Three instances where such lags may be important are capital expenditure decisions, research and development decisions, and advertising expenditure decisions (particularly where advertising is tied to new product marketing). Accordingly, firms with large capital expenditures (relative to sales), with large *R&D* expenditures (relative to sales), and with large advertising expenditure (relative to sales) will have more to gain by waiting for future information to evaluate manager performance. The corresponding hypotheses are that firms for which capital expenditures, *R&D* expenditures, or advertising expenditures are large relative to sales will be more likely to provide managers with golden parachutes.

The noise encountered in using firm performance to impute manager performance increases with the variability of nonmanager (exogenous) influences on firm performance. Taking the rate of return earned by shareholders, r, as the measure of firm performance, the greater the exogenous variation in r the more difficult (less precise) it is to impute manager performance. A measure of the noise encountered when imputing manager performance, then, would be σ_r^2, the variance of the rate of return earned by shareholders. Some variation in r for any firm i, however, can be explained by market (business cycle) factors and so this variation need not introduce noise into the imputation of manager performance. Letting r_m be the market rate of return, a time-series regression

$$r_i = \beta_0 + \beta_1 r_m + \varepsilon_i \tag{10}$$

can be used to explain some of the variation in r_i by market factors. Consequently, an alternative measure of the noise encountered when using r_i to impute manager performance would be the residual variation in (10) or the mean squared error. An hypothesis, then, is that firms for which σ_r^2 or alternatively the mean squared error for (10) is larger will be more likely to provide golden parachutes to managers.

To test the above hypotheses, a random sample of 400 firms in Standard and Poor's COMPUSTAT data base was drawn. Firms for which 1982 proxy statements were not available from Q-Data Corporation's Q-File or firms that were not U.S. corporations were

deleted from the sample. This left 331 firms. The 1982 proxy statements of these firms were examined for evidence of golden parachutes.[16] Any arrangement awarding managers compensation in the event of a "change in control" of the firm was counted as a golden parachute. Forty-seven firms were found to provide golden parachutes. The qualitative variable *Golden Parachute* was constructed with a value of 1 for firms with golden parachutes and 0 for other firms.

The COMPUSTAT data base was then used to construct additional variables for each firm. *Assets* is simply the 1982 value of the firm's assets (in millions of dollars). *Price-Earnings Ratio* is calculated by taking the year-end closing price for the firm's shares and dividing by per share earnings for each year 1980–82 and then averaging these.[17] *Capital Expenditures* is the firm's capital expenditure divided by sales again calculated for 1980–1982 and averaged. Because information on *R&D* expenditure and advertising expenditure was not reported for a substantial number of firms, the more inclusive COMPUSTAT data item entitled Selling, General, and Administrative Expenses was used to construct a measure of expenditures (other than capital expenditures) where lags between decisions and outcomes may be important. This data item includes not only advertising expenditures and *R&D* expenditures, but also foreign currency adjustments, marketing expenditures, strike expense, exploration expenditures of extractive firms and other expenditures not directly related to production. The variable *Advertising, Research, and Related Expenditures* is the firm's selling, general, and administrative expenditures divided by sales calculated for 1980–82 and averaged.

The quarterly rate of return to the firm's shareholders was calculated for each quarter 1978–82 (19 quarters) by summing dividends paid and change in share price during the quarter and dividing by share price at the beginning of the quarter. Two variables were constructed from these quarterly rates of return. The first, *Return Variability*, is simply the variance of the firm's quarterly rate of return to shareholders over the period. The second, *Unexplained Return Variability*, is the mean squared error from estimating equation (10) for each firm, using the nineteen quarterly rates of return and letting r_m be the quarterly rate of return on the Standard and Poor 400 index. Finally, information from the 1982 proxy statements was used to construct two measures of manager shareholding. The first, *Fraction Manager Owned*, is the fraction of common stock owned by the manager (defined as the highest paid employee also on the board of directors). The second, *Fraction Manager and Board Owned*, is the fraction of common stock owned by all officers and directors combined.

The empirical model with hypothesized signs of effects is

$$\text{(11)}\quad \textit{Golden Parachute} = f(\textit{Assets}(?),\ \textit{Assets}^2(?),\ \textit{Price-Earnings Ratio}(?),\ \textit{Price-Earnings Ratio}^2(?),\ \textit{Capital Expenditures}(+),\ \textit{Advertising Research and Related Expenditures}(+),\ \textit{Return Variability or Unexplained Return Variability}(+),\ \textit{Fraction Manager Owned or Fraction Manager and Board Owned}(-)),$$

where the two return variability variables are alternative measures of the noise encountered when using firm performance to impute manager performance and the two ownership variables are alternate measures of manager shareholding. The squared values of *Assets* and the *Price-Earnings Ratio* are included to

[16] Proxy statements must inform shareholders of the existing compensation arrangements with top management. Golden parachutes are a feature of such arrangements and so will be described if they exist.

[17] Firms with price-earnings ratios < 0 or > 100 were deleted.

TABLE 1—LOGIT ESTIMATES OF DETERMINANTS OF INCIDENCE OF GOLDEN PARACHUTES: DERIVATIVES OF PROBABILITY OF GOLDEN PARACHUTE EVALUATED AT VARIABLE MEAN[a]

	Variable Means				
Assets	826.220	$.644\times10^{-4}$ (1.833)	$.667\times10^{-4}$ (1.839)	$.473\times10^{-4}$ (1.546)	$.476\times10^{-4}$ (1.555)
Assets^2	312.749×10^{4}	$-.942\times10^{-8}$ (1.485)	$-.946\times10^{-8}$ (1.488)	$-.724\times10^{-8}$ (1.404)	$-.728\times10^{-8}$ (1.409)
Price-Earnings Ratio	13.347	$-.900\times10^{-2}$ (1.698)	$-.899\times10^{-2}$ (1.696)	$-.807\times10^{-2}$ (1.796)	$-.806\times10^{-2}$ (1.792)
Price-Earnings Ratio^2	303.893	$.148\times10^{-3}$ (1.999)	1.147×10^{-3} (1.997)	1.146×10^{-3} (2.170)	$.146\times10^{-3}$ (2.167)
Capital Expenditures	.108	.170 (1.695)	.171 (1.718)	.120 (1.512)	.121 (1.540)
Advertising Research, and Related Expenditure	.174	.182 (.977)	.180 (.970)	.177 (1.116)	.176 (1.109)
Return Variability	.102	.071 (.675)		.083 (.972)	
Unexplained Return Variability	.095		.071 (.662)		.082 (.948)
Fraction Manager Owned	.060	−.918 (1.836)	−.917 (1.835)		
Fraction Manager and Board Owned				−.676 (2.975)	−.675 (2.973)
Number of Observations	244	244	244	246	246
Number of Golden Parachutes	40	40	40	40	40
ln likelihood		−97.403	−97.412	−92.196	−92.222
χ^2		11.887	11.869	11.917	11.865

[a]Asymptotic *t*-ratios of logit coefficients are shown in parentheses.

allow for possible nonlinear effects of these variables. Data were not available for each firm for each variable and so the number of observations in the estimations was 244 or 246. Estimations of the four model specifications are reported in Table 1. The signs on each variable are as predicted, but significance levels are not high.[18] Table 1 provides modest support for the view of golden parachutes developed here. To determine if this view or the alternative which holds that golden parachutes are a device designed by management to deflect the discipline imposed by the market for corporate control is better supported, note that the alternative implies that golden parachutes are more likely where hostile tender offers are more likely. Consequently, the *Assets*, *Price-Earnings Ratio*, and *Fraction Manager Owned* variables would still be predicted to affect the likelihood of a golden parachute. Increases in each should make a golden parachute less likely. Under this alternative view, though, no other variables in (11) would be predicted to affect the likelihood of a golden parachute. To evaluate the merit of the two views, a likelihood ratio test was performed. First, the specifications in Table 1 were reestimated restricting all coefficients except those on the *Assets*, *Price-Earnings Ratio*, and *Fraction Manager Owned* variables to be zero. A likelihood ratio statistic, −2 (ln likelihood restricted model − ln likelihood unrestricted

[18]Since golden parachutes and shark repellents are substitutes, stronger results would be expected if firms that had either golden parachutes or shark repellents were treated alike (each given a 1 for purposes of the logit estimations of equation (11)). This was not possible because it could not be determined which firms had adopted shark repellents.

model), was then calculated. This statistic is distributed χ^2 with degrees of freedom equal to the number of restricted coefficients. It is reported at the bottom of each column in Table 1. Each is significant at just under the 5 percent level. The full model performs better than the restricted model. The incidence of golden parachutes is better explained by the view developed in this paper than that which holds they are detrimental to shareholders.

Another implication of the arguments of the previous two sections is that the timing of manager compensation will differ between firms providing and firms not providing golden parachutes. If golden parachutes arise to provide assurance that implicit deferred compensation contracts will not be subject to tender related opportunism, such compensation should be more prevalent for managers of golden parachute firms. An indication of this is a more steeply rising (at least initially) compensation-tenure profile for managers of golden parachute firms. The reason is that where compensation is delayed until better information on manager performance is available, observed compensation will increase with tenure partly because deferred compensation is being paid. This effect will become less important and may disappear as tenure increases.

For example, consider two equally productive managers. Manager *A*'s performance can be observed perfectly this period and he is paid his whole product in the current period. Manager *B*'s performance can be estimated this period but is observed perfectly one period later and so he is paid one-half his estimated product this period and one-half is delayed to be paid next period. If the actual product of both managers is \$100,000 in the first period and productivity increases with tenure at a rate of 10 percent each period, then period one compensation will be \$100,000 for Manager *A* and \$50,000 for Manager *B*. In the second period, Manager *A* will receive \$110,000 (a 10 percent increase) while Manager *B* will receive \$105,000 (a 110 percent increase). In the third period, Manager *A* will receive \$121,000 (again a 10 percent increase) and Manager *B* will receive \$115,500 (also a 10 percent increase). So the compensation-tenure profile initially rises more steeply for Manager *B*, but eventually rises at the same rate for both Managers *A* and *B*.

An hypothesis, then, is that compensation will increase more rapidly with tenure for managers of golden parachute firms, but the differential rate of increase will decline as tenure increases. To test this hypothesis, all 1982 compensation of the highest paid manager also on the board of directors was determined from the 1982 proxy statements for each of the 331 firms described previously.[19] This variable is called *Compensation*. The tenure of each manager was also determined from the proxy statements as the number of years of service on the board of directors. Since the hypothesis suggests that the effect of tenure on *Compensation* will differ between golden parachute firms and other firms, a variable *Tenure1* was defined equal to manager tenure for golden parachute firms and zero for other firms. Similarly, *Tenure0* was defined equal to manager tenure for other firms and zero for golden parachute firms. To control for other determinants of manager productivity (and so compensation), two other variables were calculated. These are manager age, *Age*, from the proxy statements and firm size, *Assets*, from COMPUSTAT. Both are expected to be positively related to manager compensation. Complete data were available for 316 of the firms.

An empirical model was estimated using the *Tenure*, *Age*, and *Asset* variables and their squared values to explain the logarithm of *Compensation*. The important predictions are that the coefficient on *Tenure1* will be larger than that on *Tenure0* and that the

[19]All compensation reported in the required Management Remuneration table of the proxy statement was summed to make this calculation. Such compensation includes salary, bonus, cash equivalent benefits, and contingent remuneration. Note that compensation which is *explicitly* deferred, and so represents an explicit legal obligation for the firm, is included in these reported figures even though it is actually paid at a later date. However, implicit deferred compensation which creates no explicit legal obligation for the firm is not included.

coefficient on $Tenure1^2$ will be smaller than that on $Tenure0^2$. The least squares estimate of the model (t-statistics in parentheses) is

$$\ln Compensation = \underset{(9.849)}{11.495} + \underset{(-.094)}{-.001}\, Tenure0$$

$$+ \underset{(2.726)}{.059}\, Tenure1 + \underset{(.292)}{.931\times 10^{-4}}\, Tenure0^2$$

$$\underset{(-2.014)}{-.001}\, Tenure1^2 + \underset{(.625)}{.028}\, Age$$

$$\underset{(-.492)}{-.199\times 10^{-3}}\, Age^2$$

$$+ \underset{(5.377)}{.600\times 10^{-4}}\, Assets \quad \underset{(-4.577)}{-.500\times 10^{-9}}\, Assets^2$$

$$R^2 = .141,\ F\text{-Statistic} = 6.307;\ N = 316.$$

The coefficient on *Tenure1* is indeed larger than that on *Tenure0*, and the coefficient of $Tenure1^2$ is smaller than that on $Tenure0^2$. A test for equality of the coefficients on the two tenure variables has an F value of 10.509 with one degree of freedom. Equality can be rejected at a significance level of less than 1 percent. A similar test on the coefficients of the squared tenure variables has an F value of 4.843 with one degree of freedom. Here, equality of the coefficients can be rejected at a significance level of 3 percent. As with the previous test on the incidence of golden parachutes, the evidence from this test on the compensation-tenure relationship provides support (albeit mild) for the hypothesis that golden parachutes are beneficial and designed to provide assurance to managers against tender related opportunism thereby enlarging the contracting possibilities between manager and shareholders.

IV

This paper develops a framework in which golden parachutes and shark repellents may be advantageous. Where contracting between managers and shareholders is an important source of discipline on manager behavior and where the best contract is implicit with much compensation delayed until manager performance can be better evaluated, managers may fear that a hostile tender offer will provide shareholders with an avenue for opportunism. A successful tender will allow the new owners to displace management and capture the delayed compensation due them. Golden parachutes and shark repellents reduce the likelihood that this will happen and so can be viewed as mechanisms adopted by shareholders to assure managers of their reliability and so induce them to accept a contract which makes both better off. It may also be true, though, that golden parachutes and shark repellents impede the disciplinary effect imposed on managers by the market for corporate control. This is the basis for the common view that they are the result of self-serving behavior by managers seeking to avoid discipline of their behavior.

The empirical part of the paper examines the incidence of golden parachutes and the differential compensation-tenure relationship for managers of golden parachute firms in a sample of 331 firms. The evidence provides some support for the view that golden parachutes are advantageous and arise to assure managers against tender related opportunism. While this evidence is not strong, it may give pause to those who would restrict the use of golden parachutes (and by analogy, shark repellents) and so limit the scope for mutually advantageous exchange between manager and shareholders.

REFERENCES

Alchian, Armen A., "Specificity, Specialization, and Coalitions," *Zeitschrift fur die gesamte Staatswissenschaft*, March 1984, *140*, 34–49.

Austin, Douglas V., "Tender Offer Update 1978–79," *Mergers and Acquisitions*, No. 2, 1980, *15*, 13–24.

______ **and Jackson, Michael J.**, "Tender Offer Update: 1984," *Mergers and Acquisitions*, No. 1, 1984, *19*, 60–69.

Becker, Gary S. and Stigler, George J., "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, January 1974, *3*, 1–18.

Bull, Clive, "The Existence of Self-Enforcing Implicit Contracts," Research Report No.

83–22, C. V. Starr Center for Applied Economics, New York University, 1983.

Carmichael, H. Lorne, "Reputations in the Labor Market," *American Economic Review*, September 1984, *74*, 713–25.

Davey, Patrick J., *Defenses Against Unnegotiated Cash Tender Offers*, Conference Board Report No. 726, 1977.

DeAngelo, Harry and Rice, Edward M., "Antitakeover Charter Amendments and Stockholder Wealth," *Journal of Financial Economics*, April 1983, *11*, 329–60.

Demsetz, Harold, "The Structure of Ownership and the Theory of the Firm," *Journal of Law and Economics*, June 1983, *26*, 375–90.

Easterbrook, Frank H. and Fischel, Daniel R., "The Proper Role of a Target's Management in Responding to a Tender Offer," *Harvard Law Review*, April 1981, *94*, 1161–204.

Eaton, Jonathan and Rosen, Harvey S., "Agency, Delayed Compensation and the Structure of Executive Remuneration," *Journal of Finance*, December 1983, *38*, 1489–505.

Fama, Eugene F., "Agency Problems and the Theory of the Firm," *Journal of Political Economy*, April 1980, *88*, 288–307.

Gilson, Ronald J., "The Case Against Shark Repellent Amendments: Structural Limitations of the Enabling Concept," *Stanford Law Review*, April 1982, *34*, 775–836.

Harris, Robert S. et al., "Characteristics of Acquired Firms: Fixed and Random Coefficient Probit Analyses," *Southern Economic Journal*, July 1982, *49*, 164–84.

Holmstrom, Bengt, "Equilibrium Long-Term Labor Contracts," *Quarterly Journal of Economics*, Suppl., 1983, *98*, 23–54.

Jarrell, Gregg A. and Bradley, Michael, "The Economic Effects of Federal and State Regulations of Cash Tender Offers," *Journal of Law and Economics*, October 1980, *23*, 371–407.

Jensen, Michael C. and Meckling, William H. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, October 1976, *3*, 305–60.

Klein, Benjamin and Leffler, Keith B., "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, October 1981, *89*, 615–41.

Lambert, Richard A., "Long-Term Contracts and Moral Hazard," *Bell Journal of Economics*, Autumn 1983, *14*, 441–52.

______ **and Larker, David F.,** "Golden Parachutes, Executive Decision-Making and Shareholder Wealth," *Journal of Accounting and Economics*, April 1985, *7*, 179–203.

Lazear, Edward, P., "Why Is There Mandatory Retirement?," *Journal of Political Economy*, December 1979, *87*, 1261–284.

Linn, Scott C. and McConnell, John J., "An Empirical Investigation of the Impact of 'Antitakeover' Amendments on Common Stock Prices," *Journal of Financial Economics*, April 1983, *11*, 361–99.

Lowenstein, Lewis, "Pruning Deadwood in Hostile Takeovers: A Proposal for Legislation," *Columbia Law Review*, March 1983, *83*, 249–334.

Manne, Henry G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, April 1965, *73*, 110–120.

Morrison, Ann M., "Those Executive Bailout Deals," *Fortune*, December 13, 1982, 82–87.

Schwartz, Steven, "Factors Affecting the Probability of Being Acquired: Evidence for the United States," *Economic Journal*, June 1982, *92*, 391–98.

Telser, Lester G., "A Theory of Self-Enforcing Agreements," *Journal of Business*, January 1980, *53*, 27–44.

Q-Data Corporation, Q-File, 1983.

Standard and Poor's Compustat Services, COMPUSTAT, 1983.

Wall Street Journal, "SEC Endorses Major Changes in Merger Fights," March 14, 1984.

______, "GM and Ford Bonuses Raise Questions about Import Curbs, Union's Restraint," April 16, 1984, p. 8.

Inventories and Interest Rates: A Critique of the Buffer Stock Model

By David G. Bivin*

Economists have devoted a great deal of effort in recent years toward demonstrating that inventories are inversely related to the interest rate.[1] Two standard assumptions of these studies are that firms possess a target level of stocks that increases with expected sales and decreases with carrying cost, and that inventories are only gradually adjusted toward the target stock. The result is the following form of the flexible accelerator-buffer stock model:

$$(1)\quad H_t - H_{t-1} = \theta(H_t^* - H_{t-1}),$$

$$(2)\quad H_t^* = \alpha + \beta\hat{S}_t + \gamma\delta_t,$$

$$0 < \theta < 1, \quad \beta > 0, \quad \gamma > 0,$$

where H_t is inventories at the end of period t, S_t is sales during period t, and δ_t is the rate of return on a unit of inventory during period t (equal to the negative of carrying cost). The asterisk and circumflex denote target and expected quantities, respectively.

The model contained in (1) and (2) is not complete until δ_t has been defined. The theory does not indicate an appropriate measure of the rate of return, but it is generally assumed that δ_t is equal to the difference between the rate of inflation on the firm's output minus the opportunity cost of funds.[2] This expression is derived as the rate of return that equates the revenue from selling a unit of output in the current period with the anticipated revenue from selling the unit in the following period less the opportunity cost of delaying the sale:

$$(3)\quad P_t = ((\hat{P}_{t+1} - R_t P_t)/(1+\delta_t))$$

$$\delta_t = \hat{\pi}_t - R_t$$

where P_t is the price of the firm's output, R_t is the short-term interest rate, and $\hat{\pi}_t = (\hat{P}_{t+1} - P_t)/P_t$ is the expected rate of inflation on the firm's output, all at time t.

This paper takes exception to this formulation of the inventory investment function on two counts. The first is that δ_t is an inadequate measure of the rate of return on finished good inventories, and the second is that the specification of desired stocks is internally inconsistent. These criticisms are developed in Section I; Section II presents an alternative model that overcomes these objections; and in Section III the model is tested. The model contains numerous implications, virtually all of which are consistent with the evidence.

I. Critique of the Buffer Stock Model

Since finished good inventories enable the firm to determine sales and output independently, both demand and supply considerations should be incorporated into models of finished good inventory investment.[3] In par-

*Assistant Professor of Economics, Indiana University, Indianapolis, IN 46202. I thank Michael R. Edgmand, John A. Carlson, Ronald L. Moomaw, John D. Rea, and an anonymous referee for comments.

[1] For recent examples of inventory models that incorporate interest rate effects, see F. Owen Irvine (1981a,b), Laura Rubin (1979–80), Charles Lieberman (1980), Dan Bechter and Stephen Pollock (1980), Louis Maccini and Robert Rossana (1981, 1984), Alan Blinder (1981, 1984), and M. A. Akhtar (1983).

[2] In some instances, these variables are entered separately in order to capture some specific effect. Akhtar, for example, regresses total manufacturing stocks on both the inflation and the interest rate in order to account for differences in tax treatment arising from *LIFO* vs. *FIFO* accounting and other factors.

[3] The importance of incorporating costs into the inventory investment function has been emphasized by John Carlson (1984), Maccini and Rossana (1984), and Blinder (1984).

ticular, inventory accumulation in the current period requires not only that it be profitable to delay sales until the following period, but also that it be profitable to produce the output in the current period. The standard specification of the rate of return captures the sales decision, but it omits the rate of return for the decision to supply output in the current period. This rate of return is found from

$$(4)\quad v_t + (R_t v_t/(1+\sigma_t)) = \hat{v}_{t+1}/(1+\sigma_t),$$

where v_t is the marginal cost of production at time t and σ_t is the internal rate of return on the decision to produce output in the current, rather than forthcoming, period. The expression on the left-hand side is the current cost of production plus discounted financing cost, while $\hat{v}_{t+1}(1+\sigma_t)^{-1}$ is the discounted expected marginal cost of production in the following period. Rearranging (4) yields

$$(5)\qquad \sigma_t = \hat{\pi}_{vt} - R_t,$$

where $\hat{\pi}_{vt} = (\hat{v}_{t+1} - v_t)/v_t$ is the expected rate of (marginal) cost inflation during period t. An increase in σ_t indicates an increase in the profitability of producing output now for future sale.

An increase in δ_t and/or σ_t results in an increase in finished good inventory investment (*ceteris paribus*). By the same token, the effect of a change in δ_t can be offset by a change in σ_t in the opposite direction. For instance, if price inflation is increasing and cost inflation is declining, the firm will delay production in order to avoid the carrying cost. Thus finished good inventories might remain constant or even fall. This trend is reversed when the rate at which costs are increasing begins to rise.

A second drawback to the buffer stock model is that it does not identify the means by which a change in the interest rate brings about a change in inventories. The presence of δ_t (and absence of σ_t) in most models suggests that interest rates influence inventories solely through changes in sales. This line of causality requires that sales be endogenous because, by definition, if sales are exogenous they are beyond the influence of the firm.[4] But if sales are endogenous, then (a) the concept of *expected* sales is not meaningful, (b) the effect of δ_t on inventories is fully accounted for by changes in planned sales, and (c) least squares estimates are subject to simultaneity bias.[5] Regardless of whether sales are exogenous or endogenous, the model is misspecified; either it includes the rate of return on a decision the firm does not have the power to make, or the effects of that decision are captured by other explanatory variables.

The shortcomings of the buffer stock model are due to the absence of a well-defined structural system derived from a consistent set of assumptions.[6] As a result, the analyst cannot ascertain whether the assumed decision rules reflect optimization. Instead, an alternative motivation (such as the buffer stock or production smoothing motive) is arbitrarily provided. This leads to a situation in which models of inventory investment are accepted or rejected on the basis of *ad hoc* criteria. Popular examples of such criteria are high estimated speed of adjustment coefficients (θ) and significantly negative coefficients on carrying cost. Neither of these results are necessarily consequences of optimization and may be contrary to the prediction obtained from a consistent set of structural equations.

II. An Alternative Model

The preferable approach is that employed by Charles Holt et al. (1960), Gerald Childs (1967), David Belsley (1969), George Hay

[4] The possibility exists that sales, although exogenous, are nevertheless correlated with R_t. But this is a statement about the firm's demand curve rather than the inventory response to a change in δ_t which economists have been trying to identify.

[5] John Gould (1969) discusses the difficulties that arise when the target stock is defined as a function of endogenous variables for the case of fixed capital.

[6] See Carlson and William Wehrs (1974) for a more complete discussion of the methodology of the flexible accelerator-buffer stock model and a derivation of joint restrictions on the parameters of the model. Martin Feldstein and Alan Auerbach (1976) present an alternative model that, in essence, reinterprets the estimated models in such a way as to increase their intuitive appeal.

(1970), and Alan Blinder (1982, 1984). These authors begin with a definition of cost or profit, and derive their decision rules from this definition and the assumption of optimizing behavior.

As an example, consider a firm, producing solely to stock, described by the following objective function:

$$(6)\quad \max_{X,H,P} \Pi = \sum_{k=0}^{T} \lambda^{k+1}\Big\{ P_{t+k}(\hat{A}_{t+k} - fP_{t+k}) - \big[aX_{t+k}^2 + \hat{v}_{t+k}X_{t+k} + b(X_{t+k} - X_{t+k-1})^2 + c(H_{t+k} - H^*)^2 + d(P_{t+k} - P_{t+k-1})^2 + \hat{C}_{t+k}H_{t+k}\big]\Big\}$$

subject to

$$(7)\qquad H_t - H_{t-1} = X_t - (\hat{A}_t - fP_t)$$

where P_t = price at time t, X_t = output at time t, H_t = finished good inventories at the end of time t, $\hat{v}_t$ = expected material and labor cost per unit of output at time t, $\hat{A}_t$ = expected level of new orders (or equivalently, sales) when the firm's price is zero at time t ($\hat{A}_t - fP_t$ is the anticipated level of new orders in period t), $\hat{C}_t$ = expected financing cost per unit of finished good inventories at time t, λ = constant discount factor, and a, b, c, d, f are constant and nonnegative.

All expectations are formed at the beginning of period t. The expression in braces is the planned level of profit in period $t+k$ contingent upon the anticipated values of A_{t+k}, v_{t+k}, and C_{t+k}. The expression in brackets is a fairly standard cost function. Profit is defined as a quadratic function of the endogenous and exogenous variables in order to yield linear decision rules and to insure the applicability of theorems on certainty equivalence developed by Herbert Simon (1956) and Henri Theil (1964). This quadratic structure requires that the discount factor be constant over time. It is assumed that revenue is received and costs are paid at the end of the period.

The restrictions on a, b, c, and d insure constant or increasing cost associated with output, inventory storage, and the magnitude of the single-period change in output and price. The level of inventories which minimizes storage cost (H^*) is assumed positive and constant.[7] The implications of this assumption for inventory behavior in the short and long run are discussed below. Profit is defined in real terms, thus P_{t+k}, $\hat{v}_{t+k}$, and $\hat{C}_{t+k}$ are expressed as ratios of their respective nominal values to the deflator and the parameters a, b, c, and d are assumed constant.

The firm described by (6) selects output and price at the beginning of the period conditional upon expected levels of demand and costs over the time horizon as well as the inherited values of inventories, output, and price. Over the course of the period, the firm may gain insights into the true value of the anticipated variables and, to the extent that production and price are flexible, will adjust accordingly. The true value of demand is realized at the end of the period. The difference between output and sales then determines the level of inventory investment.

Financing cost in the current period is the product of the nominal interest rate and the real value of stocks held during the period. The quantity of stocks held during the period is determined by the level of output and price in period $t-1$. Thus current finance cost is predetermined and may be ignored. But the price and output decisions made in the current period will determine the quantity of stocks held for the duration of the following period. This cost is

$$C_{t+k} = (h_{t+k+1}R_{t+k+1})/(1 + R_{t+k+1})$$

where h_t is the per unit real value of invento-

[7]The assumption that H^* is positive is an *ad hoc* device for rationalizing the observation that firms generally hold positive amounts of finished good inventories. This type of function has been previously employed by Belsley, Blinder (1982, 1984), and others. The assumption that H^* is constant is made for the sake of simplicity. A common alternative is that H^* is an increasing function of $\hat{A}_t$. This possibility is explored in the interpretations of the regression results.

ries and R_t is the nominal interest rate. Because this cost is not incurred until the end of the following period, it is discounted relative to revenue and other costs.

The inventory investment function implied by profit maximization for the case of a two-period time horizon is

$$(8)\quad H_t = \alpha_0 + \alpha_1 \hat{A}_t + \alpha_2 \hat{v}_t + \alpha_3 \hat{C}_t + \alpha_4 \hat{A}_{t+1} + \alpha_5 \hat{v}_{t+1} + \alpha_6 \hat{C}_{t+1} + \alpha_7 H_{t-1} + \alpha_8 X_{t-1} + \alpha_9 P_{t-1}.$$

The parameter space restrictions implied by the form of the model and the restriction that a, b, c, d, and f all be nonnegative are presented in the Appendix.[8]

This specification differs from the models criticized above in several respects. First, only the exogenous part of sales enters the decision rules. Since price is an endogenous variable, δ_t is determined as a by-product of profit maximization. The appropriate measure of financing cost per real dollar of inventory is the (discounted) nominal short-term rate. Second, the firm's inventory investment function is derived as an implication of profit maximization rather than some arbitrarily defined motive. Alternatively, the model offers a variety of reasons for the firm to maintain stocks including production smoothing ($b > 0$), price smoothing ($d > 0$), and speculative motives. Moreover, the procedure through which (8) is derived may also be used to obtain decision rules for output and price which are consistent with (8) and with profit maximization. Third, a multiperiod time horizon (along with appropriate parameter space restrictions) is easily incorporated into the model. Finally, inventories do not exhibit buffer stock behavior. That is, a permanent increase in demand results in a permanent reduction in finished good inventories. This is because of the reliance on increasing marginal cost and the assumption that H^* is constant rather than an increasing function of $\hat{A}_t$.

III. Empirical Results

The decision rules for finished good inventories, output, and price are estimated using monthly seasonally adjusted observations on the food and kindred product industry, SIC 20.[9] This is the largest 2-digit SIC production-to-stock industry. The model is estimated over the February 1968 through December 1983 period with industry-specific data provided by the Bureau of Economic Analysis (BEA).

Inventories by stage of fabrication and their respective deflators are described in John Hinrichs and Anthony Eckman (1981). The "inventory acquisition cost index" they constructed for finished goods measures the per unit cost of incrementing finished good inventories at current factor prices.[10] Thus it provides a good proxy for nominal per unit production cost (v_t). This index is also used as a proxy for h_t.[11] The interest rate necessary to complete the definition of C_t is the six-month commercial paper rate. The BEA also provides nominal and real shipments data; the ratio of these two series is the shipments deflator (P_t). Output is obtained from the standard accounting identity, $X_t = Q_t + (H_t - H_{t-1})$, where Q_t is real shipments.

Demand is assumed to be a linear function of own-price (P_t), the producer's price index for crude materials for further processing

[8] Imposing parameter space restrictions by intuition at this point largely defeats the purpose of constructing the model. Belsley, for instance, constructs a model that is similar to that employed here in that the minimum cost level of stocks is independent of expected sales. He then imposes restrictions on his parameters designed to yield buffer stock behavior. These restrictions directly contradict the original specification.

[9] An alternative estimation strategy has been applied to the automobile industry by Olivier Blanchard (1983).

[10] Kenneth West (1983) has pointed out that inventories are deflated on a cost basis while sales are deflated on a market value basis. Thus one dollar of inventories represents a greater quantity than one dollar of sales. To correct for this, the finished good inventory data obtained from the BEA have been multiplied by a constant adjustment factor constructed by West for the food and kindred product industry.

[11] These measures performed as well as several alternatives and yielded similar results.

$(\bar{P}_t)$, food wholesales (WS_t), the short-term interest rate (R_t), and lagged values of new orders. The food wholesale series is provided in real terms on a monthly basis by the BEA. The demand function is estimated using the instrumental variables approach with P_t estimated as a function of current and lagged values of v_t and lagged values of price and sales. This substitution is necessary because it is assumed below that the sales price depends partly on forecast errors on demand. (All price and cost indices are deflated by the Consumer Price Index and are expressed on a scale of approximately 1000. The interest rate is expressed as a monthly percentage rate.)

$$(9)\quad \underset{(4.5)}{Q_t=1776.629}-\underset{(2.1)}{.812P_t}+\underset{(1.9)}{.505\bar{P}_t}-\underset{(2.6)}{171.612R_t}+\underset{(7.3)}{.491Q_{t-1}}+\underset{(3.8)}{.247Q_{t-2}}+\underset{(4.8)}{.136WS_t}$$

$$\text{Durbin } h=-1.27;\quad \bar{R}^2=.98.$$

The forecast value of A_t is

$$(10)\quad \hat{A}_t=1776.629+.505\bar{P}_t-171.612R_t+.491Q_{t-1}+.247Q_{t-2}+.136\hat{W}S_t.$$

The values of R_t and $\bar{P}_t$ are assumed known at the beginning of the period, but the actual and predicted values of A_t will differ because of the random error, and because the forecast of wholesales is not perfect. The firm is also assumed to know v_t at the beginning of the period, but C_t must be forecast since it involves values that are not observed until the following month. Following Blinder (1984), all forecasts (except for $\hat{A}_t$) are formulated as 12th-order autoregressive processes. Forecasts of values occurring in period $t+1$ are obtained using the chain rule except for $\hat{A}_{t+1}$ which is obtained in an analogous manner to $\hat{A}_t$. The values of $\hat{A}_{t+1}$, $\hat{C}_{t+1}$, and $\hat{w}_{t+1}$ collectively (and individually) provided no significant explanatory power in any of the models, and so a single-period horizon was assumed.

The model was originally estimated using ordinary least squares. This procedure implicitly incorporates the accounting identity that defines finished good inventory investment. There was evidence of autocorrelation in the price equation. A maximum likelihood, grid-search procedure was applied to this equation with a maximum of the log likelihood function achieved at a value of *rho* equal to one. Unfortunately, the serial correlation remained. Some experimentation indicated that including v_{t-1} as an explanatory variable yielded random residuals and a maximum of the log likelihood function at a value of *rho* of about .4. Thus v_t and v_{t-1} are included together as a measure of current production cost.

The modified equations are estimated jointly using a nonlinear maximum likelihood procedure that insures that the estimates will satisfy the accounting identity. This requires that the lagged residual from the price equation appear in the output and inventory equations. The estimated coefficients are presented in Table 1 along with the Durbin h, the standard error of the regression, and the coefficient on the lagged value of the residual from the price equation (ε^P_{t-1}).[12] The forecasting error on current new orders, e_t, is included in the model in

[12]Sensitivity analysis indicated the existence of dual solutions to the maximum likelihood problem. The first is characterized by a value of *rho* from the price equation of about 0.4 and implies a slow rate of price adjustment. The second is characterized by a value of *rho* in the vicinity of one and implies a much more rapid speed of adjustment. The "low *rho*" solution is preferable because it yields a slightly lower value for the criterion function, a lower residual sum of squares for both the inventory and output equations, and because the results conform more closely to the theory. The "high *rho*" solution is also generally consistent with the theory. The results differ primarily in that there are offsetting changes in the coefficients of P_{t-1} and v_t+v_{t-1}. These results are available upon request.

Blinder has noted the existence of low *rho* and high *rho* solutions in models of inventory behavior. He suggests that the dual solutions arise because of the difficulty in distinguishing between partial adjustment and autocorrelation (1984, p. 33). His empirical analysis indicates that the 2-digit manufacturing industries are generally low *rho*, but his conclusion on the food and kindred product industry is sensitive to the specification of the model (p. 42).

TABLE 1—ESTIMATES OF THE DECISION RULES[a]

	H_t	X_t	P_t
Constant	150.964	116.299	42.708
	(.7)	(.5)	(1.4)
$\hat{A}_t$	−.099	.894	.009
	(1.7)	(15.6)	(2.1)
e_t	.015	1.006	.010
	(.3)	(22.2)	(3.4)
$v_t + v_{t-1}$[b]	−.290	−.379	.110
	(2.0)	(2.2)	(2.7)
C_t	−.579	−.451	−.158
	(.2)	(.1)	(.4)
H_{t-1}	.942	−.054	−.005
	(34.7)	(1.9)	(1.6)
X_{t-1}	.115	.123	−.009
	(2.0)	(2.1)	(2.4)
P_{t-1}	.477	−.241	.884
	(2.8)	(1.2)	(20.9)
ε^P_{t-1}	−2.468	−2.797	.405
	(2.4)	(2.7)	(5.2)
Durbin h	−.3	−.44	.8
Standard Error	93.190	93.390	6.562
Mean Value of the Dependent Variable	7214.	10381	1024.

[a]The t-statistics are shown in parentheses.

[b]The entries in this line are the sum of the coefficients of these two variables with the weights assigned to each variable constrained to be equal across equations.

TABLE 2—LONG-RUN RESPONSES[a]

	H	X	P
A	.194	1.011	−.014
	(1.6)	(95.6)	(1.1)
v	1.252	−.770	.949
	(.9)	(7.3)	(7.3)
C	−14.651	.584	−.720
	(.3)	(.2)	(.2)

[a]The t-statistics are shown in parentheses (calculated through linear approximation).

order to account for unintended inventory investment and inner-period price and output adjustments.

The estimates satisfy the parameter space restrictions except that the coefficient on e_t should be negative in the inventory equation. The coefficients on $\hat{C}_t$ are all insignificant, thus the estimates provide no evidence of a significant inverse relationship between finished good inventories and the interest rate in the short run.

The coefficients on the lagged variables are all of the proper sign and magnitude. Note that inventories provide both production and price smoothing. The theory predicts that an anticipated increase in demand will yield an increase in output and price, and a reduction in planned end-of-period stocks. The last conclusion is contrary to the buffer stock model and arises as a direct consequence of increasing marginal production cost. The coefficients on anticipated demand are consistent with the theory. The size of the coefficient on e_t in the output equation reflects a high degree of production flexibility in response to forecasting errors. The coefficients on current production cost are all of the proper sign. If production cost rises, the firm will reduce output and raise price. The latter can be viewed as a means of lowering sales in order to reduce the disinvestment in stocks resulting from the decline in production.

A major difficulty with assigning parameter space restrictions by simulation is that some of the more subtle (for example, cross-equation) restrictions are likely to be overlooked. Fortunately, many of the restrictions yield implications which are evident in the long run. Thus completeness requires that we compare the steady-state solutions implied by the estimates with the parameter space restrictions implied by the model (see the Appendix). Table 2 contains the estimated long-run responses of the endogenous variables to permanent unit increases in the exogenous variables.

The theory predicts an inverse long-run relationship between demand and inventories because of increasing marginal cost of production and the assumption that H^* is constant. The positive coefficient contradicts the theory, but is insignificant at the 5 percent level. The negative short-run response coupled with the positive long-run response is consistent with the notion that inventories are determined by a buffer stock motive that is dominated in the short run by costs of adjusting output and price. The results also indicate a one-for-one correspondence between changes in demand and changes in output in the long run. The magnitude of the output response and the sign of the inventory response suggest that firms do not

smooth production in the long run. This conclusion has been obtained previously by Kenneth West (1985), Olivier Blanchard, and Blinder (1984). The significant short-run price response to demand changes indicated in Table 1 is not apparent in the long run. In fact, the long-run price response should be positive.

The long-run response of finished good inventories to a change in financing cost is larger than one might expect from the coefficients in Table 1. The difference in magnitudes can be attributed to the cost of changing price and output that dampens inventory investment in the short run. The coefficients on C for price and output are supposed to be zero: as inventories adjust to their new long-run level, price and output both return to their original levels. The interest rate responses are all insignificant and highly inelastic (the elasticities for inventories, output, and price are $-.014$, $.003$ and $-.004$, respectively). The signs on production cost conform to the theory for output and price, but should be negative in the inventory equation.

IV. Conclusions

This paper presents several reasons why the buffer stock model of inventory investment is misspecified. The misspecification occurs because modelers rely upon their intuition to specify the reduced-form system rather than concentrating their efforts on the structural underpinnings. Although the models appear sound on the surface, the link between carrying cost and inventory investment is tenuous because of failures in properly defining the independent variables and in accounting for the causal route from interest rates to inventories.

The paper also presents a preferable methodology in which a model of inventory investment is derived as a consequence of optimal behavior. This procedure isn't new, but it is rarely employed, presumably because it is more difficult to implement. The principal advantage of this approach is that it yields a well-defined and internally consistent set of parameter space restrictions on the short- and long-run behavior of the endogenous variables. In contrast, the flexible accelerator-buffer stock methodology arbitrarily specifies the long-run target level of stocks. The short-run behavior is then governed by an adjustment process that assumes that firms minimize inventory maintenance costs rather than maximize profits.

The model is tested using monthly observations on the food and kindred product industry. The estimates generally support the theory. No evidence is found to support the notion that finished good inventories are sensitive to the short-term interest rate. Likewise, the data do not support the hypothesis of a short-run buffer stock motive. The long-run response of finished good inventories to an increase in demand is positive but insignificant.

Appendix

The restrictions on the reduced-form parameters are derived through simulation over a wide variety of values of the structural parameters. This procedure is necessitated by the complexity of the model which prevents a direct mapping from the structural to the reduced-form parameters.

In what follows, the reduced-form parameter on the exogenous variable z in the decision rule for y is denoted as $\pi(y,z)$. The rationale underlying these restrictions is given in the text. These restrictions are presented for the two-period horizon case, but are also applicable to the single-period case.

$$\begin{aligned}
&-1<\pi(H_t,\hat{A}_t)<0,\quad 0<\pi(X_t,\hat{A}_t)<1,\\
&\qquad \pi(P_t,\hat{A}_t)>0,\quad 0<\pi(H_t,\hat{A}_{t+1})<-\pi(H_t,\hat{A}_t),\\
&\qquad\qquad 0<\pi(X_t,\hat{A}_{t+1})<\pi(X_t,\hat{A}_t),\\
&\qquad\qquad 0<\pi(P_t,\hat{A}_{t+1})<\pi(P_t,\hat{A}_t);
\end{aligned}$$

$$\begin{aligned}
&\pi(H_t,\hat{v}_t)<0,\quad \pi(X_t,\hat{v}_t)<0,\quad \pi(P_t,\hat{v}_t)>0,\\
&\qquad 0<\pi(H_t,\hat{v}_{t+1})<-\pi(H_t,\hat{v}_t),\\
&\qquad 0<\pi(X_t,\hat{v}_{t+1})<-\pi(X_t,\hat{v}_t),\\
&\qquad 0<\pi(P_t,\hat{v}_{t+1})<\pi(P_t,\hat{v}_t);
\end{aligned}$$

$$\begin{aligned}
&\pi(H_t,\hat{C}_t)<0,\qquad \pi(X_t,\hat{C}_t)<0,\qquad \pi(P_t,\hat{C}_t)<0,\\
&\pi(H_t,\hat{C}_{t+1})<0,\quad \pi(X_t,\hat{C}_{t+1})<0,\quad \pi(P_t,\hat{C}_{t+1})<0;
\end{aligned}$$

$$0 < \pi(H_t, H_{t-1}) < 1, \qquad -1 < \pi(X_t, H_{t-1}) < 0,$$
$$\pi(P_t, H_{t-1}) < 0, \qquad 0 < \pi(H_t, X_{t-1}) < 1,$$
$$0 < \pi(X_t, X_{t-1}) < 1, \qquad \pi(P_t, X_{t-1}) < 0,$$
$$\pi(H_t, P_{t-1}) > 0, \qquad \pi(X_t, P_{t-1}) < 0,$$
$$0 < \pi(P_t, P_{t-1}) < 1.$$

The reduced-form parameters connected with each exogenous variable are also constrained by the accounting identity that defines inventory investment. As noted in the text, this constraint is imposed through maximum likelihood estimation.

The restrictions on the steady-state parameters, denoted as $\gamma(y, z)$, are as follows:

$$\gamma(H,A) < 0, \quad 0 < \gamma(X,A) < 1, \quad \gamma(P,A) > 0,$$
$$\gamma(H,v) < 0, \quad \gamma(X,v) < 0, \quad \gamma(P,v) > 0,$$
$$\gamma(H,C) < 0, \quad \gamma(X,C) = 0, \quad \gamma(P,C) = 0.$$

REFERENCES

Akhtar, M. A., "Effects of Interest Rates and Inflation on Aggregate Inventory Investment in the United States," *American Economic Review*, June 1983, *73*, 319–28.

Bechter, Dan M. and Pollock, Stephen H., "Are Inventories Sensitive to Interest Rates?," *Federal Reserve Bank of Kansas City Economic Review*, April 1980, *65*, 18–27.

Belsley, David A., *Industry Production Behavior: The Order-Stock Distinction*, Amsterdam: North-Holland, 1969.

Blanchard, Olivier, "The Production and Inventory Behavior of the American Automobile Industry," *Journal of Political Economy*, June 1983, *91*, 365–400.

Blinder, Alan S., "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, 2:1981, 443–505.

______, "Inventories and Sticky Prices: More on the Microfoundations of Macroeconomics," *American Economic Review*, June 1982, *72*, 334–48.

______, "Can the Production Smoothing Model of Inventory Behavior Be Saved?," Working Paper No. 1257, National Bureau of Economic Research, 1984.

Carlson, John A., "Stocks, Shocks, and Price-Output Dynamics," Purdue Institute Paper No. 859, Purdue University, 1984.

______ **and Wehrs, William E.**, "Aggregate Inventory Behavior," in George Horwich and Paul A. Samuelson, eds., *Trade, Stability, and Macroeconomics: Essays in Honor of Lloyd A. Metzler*, New York: Academic Press, 1974, 311–32.

Childs, Gerald L., *Unfilled Orders and Inventories: A Structural Analysis*, Amsterdam: North-Holland, 1967.

Feldstein, Martin and Auerbach, Alan, "Inventory Behavior in Durable Goods Manufacturing: The Target-Adjustment Model," *Brookings Papers on Economic Activity*, 2:1976, 351–98.

Gould, John P., "The Use of Endogenous Variables in Dynamic Models of Investment," *Quarterly Journal of Economics*, November 1969, *83*, 580–99.

Hay, George A., "Production, Price, and Inventory Theory," *American Economic Review*, September 1970, *60*, 531–545.

Hinrichs, John C. and Eckman, Anthony D., "Constant Dollar Manufacturing Inventories," *Survey of Current Business*, November 1981, *61*, 16–23.

Holt, Charles F. et al., *Planning Production, Inventories, and Work Force*, Englewood Cliffs: Prentice-Hall, 1960.

Irvine, F. Owen, Jr., (1981a) "Merchant Wholesaler Inventory Investment and the Cost of Capital," *American Economic Review Proceedings*, May 1981, *71*, 23–29.

______, (1981b) "Retail Inventory Investment and the Cost of Capital," *American Economic Review*, September 1981, *71*, 633–48.

Lieberman, Charles, "Inventory Demand and Cost of Capital Effects," *Review of Economics and Statistics*, August 1980, *62*, 348–56.

Maccini, Louis J. and Rossana, Robert J., "Investment in Finished Goods Inventories: An Analysis of Adjustment Speeds," *American Economic Review Proceedings*, May 1981, *71*, 17–22.

______ **and** ______, "Joint Production, Quasi-Fixed Factors of Production, and Investment in Finished Goods Inventories," *Journal of Money, Credit, and*

Banking, May 1984, *16*, 218–36.

Rubin, Laura S., "Aggregate Inventory Behavior: Response to Uncertainty and Interest Rates," *Journal of Post Keynesian Economics*, Winter 1979–80, *2*, 201–11.

Simon, Herbert A., "Dynamic Programming Under Uncertainty With a Quadratic Criterion Function," *Econometrica*, January 1956, *24*, 74–81.

Theil, Henri, *Optimal Decision Rules for Government and Industry*, Amsterdam: North-Holland, 1964.

West, Kenneth D., "A Note on the Econometric Use of Constant Dollar Inventory Series," *Economics Letters*, 1983, *13*, 337–41.

______, "A Variance Bounds Test of the Linear Quadratic Inventory Model," Working Paper No. 1581, National Bureau of Economic Research, 1985.

Technological Innovation, Capital Mobility, and the Product Cycle in North-South Trade

By DAVID DOLLAR*

The introduction of new products is a form of technological innovation that plays an important role in determining the pattern of international trade, particularly the pattern of trade between developed and less developed nations. Raymond Vernon (1966) has argued that most new products are introduced in the developed countries, in the United States in particular, because the markets in these nations are large and in the early stages of a product's life production needs to be located close to the market. After a product has become standardized, it is possible to produce it far from the main markets.

Paul Krugman (1979) has constructed a formal model of the product cycle in which it is taken as given that there is continuous introduction of new products in the developed region, the North; at the same time the less developed region, the South, learns in each period to produce some of the goods formerly produced only in the North.[1] The key variables in the model are the *numbers* of goods that each region produces. The model has the nice feature that for each region, an increase in the range of products that can be produced, *ceteris paribus*, translates into an improvement in the region's terms of trade. This result accords well with the historical experiences of countries such as Japan and Taiwan.

This version of the product cycle is essentially a classical trade model in which there is complete specialization: the North produces only "new" goods while the South produces only "old" goods. The cycle arises because what is a new good in one period eventually becomes an old good. A major drawback of the classical model is that, because of complete specialization, the factors of production in different regions do not compete through international trade, a key insight of the neoclassical trade model that seems especially relevant today in trade between developed and less developed countries. In the classical model, an increase in the labor force in one nation generally leads to an *increase* in wages in the other nation. Krugman's product cycle model has this feature: an increase in the labor force in the South increases the demand for the North's products while leaving the supply of new goods unaffected. Hence the North's terms of trade and wages for its workers must rise.

In this paper I construct a dynamic general equilibrium model of North-South trade that attempts to combine the product cycle approach of Vernon and Krugman with the pressures toward factor-price equalization captured by the neoclassical trade model. Following Krugman, I take it as given that there is continuous introduction of new products in the North. The innovative feature of the model is that, while continuing to abstract from the costs and details of technology transfer, I make the plausible assumption that the rate of transfer is positively related to differences in production costs in the two regions. In addition, I introduce capital as a second factor of production and assume that the movement of capital between regions takes place slowly over time. These aspects are developed in the first three sections of the paper.

The assumptions characterize a long-run equilibrium that is analyzed in Sections IV and V. In the long-run equilibrium, factor prices and the terms of trade are stable. To

*Assistant Professor, Department of Economics, University of California, Los Angeles, CA 90024. This article is based on my doctoral dissertation, written under the guidance of Pentti J. K. Kouri at New York University. I thank William Baumol, Janusz Ordover, Charles Wilson, and an anonymous referee for helpful comments.

[1]Related models of North-South trade are constructed in Ronald Findlay (1980) and Krugman (1981).

attain this stability it is necessary that the ratio of the number of goods produced in each region be constant. Since the North is continuously introducing new goods, however, this can only be the case if there is simultaneously a flow of technology to the South. For there to be an incentive for this transfer of technology, production costs must be lower in the South than in the North. Furthermore, because the return to capital in each region is equalized in the long run through capital mobility, the difference in production costs must be attributed to lower labor costs in the South. Thus the North's ability to introduce and temporarily monopolize new technology enables its workers to earn a premium over the wages paid to their counterparts in the South; this difference in labor costs in turn provides the incentive for technology to be diffused to the South, driving the product cycle.

The model illustrates how the gap in wages between North and South can be stable even while the product cycle generates constant turmoil at the micro level. However, in Section VI it is shown that relative wages in the two regions will be altered if the labor force grows more rapidly in the South than in the North. The short-run effect of an increase in the labor force in the South is to raise wages in the North by increasing demand for the North's products and improving the North's terms of trade; this is the classical result mentioned above. Over time, however, the improved terms of trade for the North (indicating an increase in the North's production costs relative to the South's) lead to a more rapid rate of diffusion of technology to the South. There is also a flow of capital to the South that reduces the capital-labor ratio in the North. At the new long-run equilibrium, real wages in the North have fallen relative to the original long-run equilibrium.

In the short run, the increase in the southern labor force affects only the demand for the North's products (and hence for its labor). Over time, however, the expanded labor force in the South affects supply conditions in the North by attracting capital and technology. In the long run, workers in the different regions compete for capital and technology. The process described by the model is consistent with recent experience. In the 1960's, the main effect that population growth and industrialization in the Third World had on the economies of the developed nations was to generate demand for the North's products. In the 1970's and 1980's, however, many northern industries have been losing out in competition with Third World suppliers, significantly reducing the demand for industrial labor in the developed countries and putting downward pressure on wages there.

I. The Static Model

The world is divided into two regions, North and South. There is a large number of goods in the system, all of which can be produced using three factors of production: capital, labor, and know-how. Capital and labor are homogeneous throughout the world. The know-how to produce each good originates in the North.

The defining characteristic of "region" is differential speeds of factor mobility between regions as compared to within regions. The know-how to produce a good diffuses instantaneously throughout the North, but only becomes available to the South with a lag. Thus, at a point in time, all goods fall into one of two categories: old goods, whose technology of production is available in both regions; or new goods that have recently been developed in the North and can only be produced there.

The movement of capital between regions also takes place slowly over time, so that at each moment the stock of capital in each region is fixed. There is no movement of labor between regions even in the long run. Within each region there is instantaneous perfect mobility of labor and capital. Finally, goods can move between regions with zero transportation costs.

Once a region has the know-how to produce a good, the know-how is freely available within the region, so that when considering costs only capital and labor need be considered as inputs into the production process. I assume that the capital and labor input into the production of each good can

be represented by the same neoclassical production function with constant returns to scale.[2]

Despite the neoclassical production function, this is essentially a classical model in which the (constant) rate of transformation between two *types* of goods is different in the two regions: the opportunity cost of a new good in terms of an old good is 1 in the North and infinite in the South.[3] This creates a basis for trade. There is only one relative price that can vary, the price of a new good in terms of an old good; with perfect competition within each region the prices of all old goods must be the same, and the prices of all new goods must be the same.

As in the classical model, when trade is allowed the price of a new good relative to an old good, p, is bounded by the different rates of transformation in the two regions, 1 and infinity. If the North produces both types of goods, p must equal 1. If, on the other hand, there is complete specialization, p will be greater than 1 and will be demand determined. Geometrically, the world supply of the aggregate output of new goods is perfectly elastic at $p=1$ until all of the North's resources are employed in the production of new goods, at which point the curve becomes perfectly inelastic, as in Figure 1. Equilibrium p will be greater than 1 if world demand for new goods intersects this supply schedule on the inelastic portion, as in Figure 1.[4]

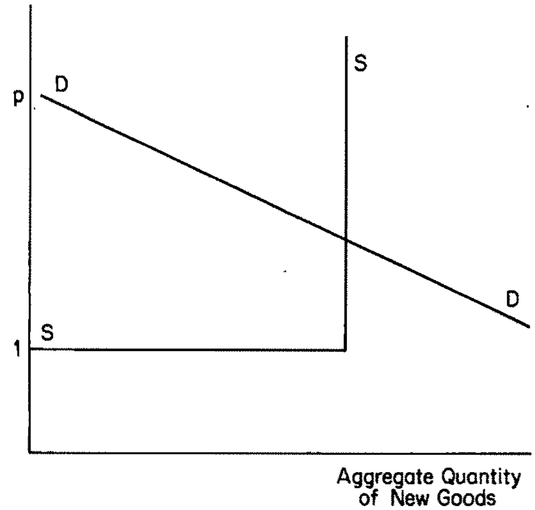


FIGURE 1

Equilibria with complete specialization are our primary interest; it is shown subsequently that if there is a static equilibrium with $p=1$ (incomplete specialization in the North), the system will move over time to an equilibrium with $p>1$. Hence I will concentrate on equilibria with complete specialization; complete specialization has the advantage that we can identify new goods with goods produced in the North and old goods with goods produced in the South. Since the numbers of new and old goods, n_N and n_S, are fixed at a point in time, the number of goods produced in the North and the number produced in the South are thus fixed at a point in time. Now we are in a position to specify formally the static equilibrium that obtains at each point in time.

A. *The Demand Side*

The demand side of the model is very simple. All individuals in both regions have the same utility function:[5]

$$U=\left(\sum^{n} c_i^{\theta}\right)^{1/\theta}, \qquad 0<\theta<1,$$

[2] The economic implication of the assumption that all goods have the same neoclassical production function is that for given factor prices the capital-labor ratio will be the same in all industries, as is assumed implicitly or explicitly by the classical economists. With perfect competition within each region, the *relative* prices of all goods produced in a region must then be constant, and it is possible to choose units so that all of these relative prices are equal to 1.

[3] The assumption that it is *impossible* to produce a new good in the South is actually unnecessarily strong. Assuming that the rate of transformation between old and new goods in the South is constant and greater than 1 preserves the essential structure of the model.

[4] Since the relative price between any two new goods is constant, we can treat new goods as a composite commodity for which it is possible to draw a single supply schedule and a single demand schedule.

[5] This utility function is borrowed from Avinash Dixit and Joseph Stiglitz (1977).

where c_i is consumption of the ith good and n is the total number of goods. The total number of products comprises both old and new goods, so that $n = n_S + n_N$. The utility function indicates that all goods enter demand symmetrically. Two goods with the same price will be consumed in the same quantity by all consumers; thus all goods produced in the North, which must have the same price, will be produced in the same quantity; all goods produced in the South will also have the same price and be produced in the same quantity. This enables us to speak of representative northern and southern goods.

It should be pointed out that the utility function has the special feature that, with a given income, a consumer's welfare will improve if the number of products to consume increases.

It remains only to determine the ratio in which all consumers will purchase new and old goods. The utility function implies that the demand for a new (northern) good relative to the demand for an old (southern) good will be a function only of the relative price of the two types of goods:

$$c_N/c_S = p^{-(1/1-\theta)}, \tag{1}$$

where c_j is world consumption of a representative good from region j.

B. *The Supply Side*

Within each region perfect competition prevails. This means that all goods produced in a region have the same price, which is equal to average cost. Choosing a southern good as numeraire, this relationship can be expressed for each region as

$$a_{LN}(w_N/q_N)w_N + a_{KN}(w_N/q_N)q_N = p \tag{2}$$

$$a_{LS}(w_S/q_S)w_S + a_{KS}(w_S/q_S)q_S = 1 \tag{3}$$

where a_{ij} is the amount of factor i used per unit of output in region j, w_j is the wage in region j, q_j is the return per unit of capital in region j, and p is the terms of trade between a northern (new) good and a southern (old) good.[6] The constant returns to scale in production mean that the factor coefficients are functions of *relative* factor prices only. Obviously the derivatives of the labor coefficients are negative, and the derivatives of the capital coefficients are positive.

The other element on the supply side is the assumption that factor prices will adjust so as to bring about full employment of resources in both regions. Equating the demand and supply of each factor yields

$$a_{LN}(w_N/q_N)n_N c_N = L_N \tag{4}$$

$$a_{KN}(w_N/q_N)n_N c_N = K_N \tag{5}$$

$$a_{LS}(w_S/q_S)n_S c_S = L_S \tag{6}$$

$$a_{KS}(w_S/q_S)n_S c_S = K_S \tag{7}$$

where L_j and K_j are the supplies of labor and capital in region j. The left-hand side of equation (4) is the demand for labor in the North: the amount of labor used per unit of output when factor prices are w_N/q_N times the total output of a representative northern good times the number of such goods; this is equated with the fixed supply of labor. Equations (5)–(7) are constructed in the same way.

Equations (1)–(7) provide seven equilibrium conditions in the variables w_N, q_N, w_S, q_S, c_N, c_S, and p.[7] In equilibrium each region exports some quantity of each of the goods that it produces. There is a flow of new goods from North to South, and a flow

[6]The labor coefficients, $a_{LN}(\cdot)$ and $a_{LS}(\cdot)$, are the same *function*, as are the capital coefficients, $a_{KN}(\cdot)$ and $a_{KS}(\cdot)$. This follows because all goods have the same production function. But since relative factor prices will in general be different in the two regions, the coefficients will have different *values* in the two regions; for this reason it is useful to employ the subscripts S and N.

[7]As long as the production technology is such that production cannot take place with only one factor of production, there must be a unique all-positive solution vector to equations (1)–(7), given positive parameter values. At this equilibrium, the Marshall-Lerner condition is satisfied.

of old goods in the opposite direction. The regions' budget constraints ensure that this trade is balanced in value terms.

II. Comparative Statics

In the static model, consumption patterns, factor prices, and the terms of trade are determined by consumer tastes, technology, the numbers of new and old goods, and the supplies of factors in each region. In the next section I introduce assumptions about how some of these parameters change continuously over time. First, however, it is necessary to investigate how the variables of the static system change as these parameters change.

To reduce the number of variables with which we are working, it is convenient to define $r = n_N/n_S$, $k = K_N/K_S$, and $q = q_N/q_S$. The variables r and k are the ratios of the number of goods produced in each region and the stocks of capital in each region; q is the ratio of the return to capital in the North to the return to capital in the South.

The variables p and q are implicit functions of the parameters of the static model. It turns out, however, that the terms of trade and factor prices are not altered by equiproportionate changes in the number of goods produced in each region, so that p and q are functions of r, but not of the absolute levels of n_N and n_S. Also, it should be obvious that K_N and K_S can be replaced by k and K, the latter being the total stock of capital in the world.

Hence the functions $p(\cdot)$ and $q(\cdot)$ can be written as

$$p = p(r, k, L_S, L_N, K, \theta),$$

$$q = q(r, k, L_S, L_N, K, \theta).$$

In the subsequent analysis the labor supply in the North, L_N, the total stock of capital K, and the demand parameter θ will be held constant; thus for simplicity these three arguments will be dropped from the functions.

The partial derivatives of p and q with respect to the remaining three arguments, r, k, and L_S, can be determined by analyzing the comparative statics around the static equilibrium.[8] These derivatives are easily understood intuitively. When r increases, *ceteris paribus*, both p and q rise. Because of the special form of the utility function, if the number of goods produced in the North increases relative to the number produced in the South, aggregate demand for the North's output increases at the existing terms of trade. Since the North's supply of output is fixed, p must rise to equilibrate the system; with factor supplies constant this increases q_N proportionally and leaves q_S unchanged. Thus the partial derivatives of p and q with respect to r satisfy $p_1/p = q_1/q > 0$. This analysis implies that either region can improve its terms of trade by extending the range of goods that is produces.[9]

If k increases, other things (including the total stock of capital in the world) held constant, aggregate output in the North (South) rises (falls), so that there is excess supply of the North's output at the existing terms of trade and p must fall to equilibrate the system. The increase in the capital-labor ratio in the North implies that q_N falls proportionally more than p while q_S rises, so that the partial derivatives of p and q with respect to k satisfy $q_2/q < p_2/p < 0$.

Finally, if L_S increases, *ceteris paribus*, the South's aggregate output increases, requiring a rise in p to equilibrate the system. This leads to a proportional increase in q_N; q_S rises as well because capital is now relatively more scarce in the South. The change in q is thus of ambiguous sign, but if positive the change must be relatively smaller than the increase in p, so that $p_3/p > q_3/q$ (where $p_3/p > 0$).

[8] These derivatives are derived explicitly in my dissertation.

[9] Innovation generally has the opposite effect when it takes the form of improvements in the production processes for existing goods. In the Ricardian model developed by Rudiger Dornbusch, Stanley Fischer, and Paul Samuelson (1977), for instance, technical progress will reduce the terms of trade for the innovating nation. For a discussion of technological progress in a Heckscher-Ohlin model, see Ronald Jones (1970).

III. Innovation, Technology Transfer, and the Flow of Capital

In the static model the number of goods produced in each region and the supplies of factors are parameters that play a key role in determining the trade equilibrium, as is generally the case in static models of trade. The innovative feature of the product cycle model is that, over time, these parameters change in a predictable way. The introduction of new products will increase the number of goods produced in the North. The transfer of technology, on the other hand, will change new goods into old goods, reducing the number of goods produced in the North and increasing the number produced in the South. The movement of capital between regions will affect the supply of capital in both regions.

In this section, I introduce assumptions about innovation, technology transfer, and the movement of capital that make the supply of capital and the number of goods produced in each region endogenous in the long run. I will continue to treat the supplies of labor as exogenous parameters, with no movement of labor between regions allowed.

The assumptions introduced in this section imply that the time derivatives of r and k are functions of q and p, which in turn are functions of r, k, and the other parameters of the static system. The end result is a system of two differential equations in which the time derivatives of r and k are functions of the values of r and k and parameters. Setting these equations equal to zero defines a long-run equilibrium in which the ratios of the number of goods produced and the supplies of capital in each region are constant and hence factor prices and the terms of trade are stable.

A. *Technological Innovation and Technology Transfer*

As noted in the introduction, I take it as given that there is continual introduction of new goods in the North. This innovation increases the total number of goods available, n. I assume that the time derivative of n is proportional to the number of new goods currently being produced in the North; the rationale for this is that innovation only takes place in the North and thus the absolute number of products introduced should be related to the size of the North's economy as reflected in n_N, the number of new goods currently being produced there. Specifically,

$$\dot{n} = in_N \tag{8}$$

where i is a positive constant.[10]

The transfer of technology enables the South to produce goods that had previously been the monopoly of the North. The incentive for transferring production to the South is that, once the South has access to the technology, it can produce any good more cheaply than it can be produced in the North. This difference in costs creates the potential for economic profits; though economic profits are never specifically realized in the model, the potential for them nevertheless provides an incentive for firms in the North to move production to the South, and for firms in the South to learn to imitate northern products.

Abstracting from the costs and details of technology transfer, I assume simply that the rate at which new goods become old goods is a positive function of the difference in the costs of production in the two regions: the greater this difference in cost, the greater are the potential economic profits to be gained through technology transfer, and hence the more such transfer takes place. The difference in the costs of production in North and South is conveniently summed up by the terms of trade, p. Since all goods can be produced with the same production function, p is the ratio of the minimum cost of production in North and South of any good that can be produced in either region. Thus as p increases, the transfer of technology should increase. Specifically, the time derivative of n_S, the number of old goods, is

$$\dot{n}_S = f(p(r,k,L_S))n_S, \qquad f' > 0, \qquad f(1) = 0. \tag{9}$$

[10] Though strictly speaking n, n_N, and n_S can only take on integer values, I will treat them as continuous variables.

This formulation means that when p is constant the number of goods produced in the South grows at a constant percentage rate. When p is equal to 1 the costs of production in the two regions are equal and there is no longer an incentive to transfer technology from North to South.

As long as complete specialization prevails, any good that the South learns to produce will cease to be produced in the North. Thus the rate of change of the number of goods produced in the North, $\dot{n}_N$, is equal to the rate of introduction of new products, $\dot{n}$, minus the rate of transfer, $\dot{n}_S$:

$$(10)\quad \dot{n}_N = in_N - f(p(r,k,L_S))n_S.$$

What we are interested in is the rate of change of $r = n_N/n_S$, which is

$$(11)\quad \dot{r} = ir - (1+r)f(p(r,k,L_S)).$$

Thus the processes of innovation and technology transfer together determine how the ratio of the number of goods produced in each region changes over time.

B. *The Movement of Capital*

For simplicity I assume that the total stock of capital in the world is fixed and indestructible; but over time capital can be moved from one region to another. This movement will take place in response to differences in the return to capital in each region. These two assumptions together imply that the time derivative of k, the ratio of the capital stock in the North to the capital stock in the South, is

$$(12)\quad \dot{k} = g(q(r,k,L_S)),\ \ g' > 0,\ \ g(1) = 0.$$

The share of the world's capital stock employed in the North (South) increases when q is greater (less) than 1. The flow of capital stops when the return to capital is equalized in the two regions.[11]

IV. Dynamic Adjustment to Long-Run Equilibrium

Equations (11) and (12) are two differential equations in two variables, r and k. They show how r and k change over time as functions of their own values and the parameter L_S. Thus the number of goods produced and the stock of capital in each region at a point in time determine factor prices and the terms of trade. The values of these variables in turn affect the processes of innovation, technology transfer, and capital movement, which result in changes in the values of r and k. As long as r and k are changing, factor prices and the terms of trade will be constantly changing.[12]

Long-run equilibrium occurs when r and k attain stable values; that is, when $\dot{r} = 0$ and $\dot{k} = 0$. We can examine the nature of this equilibrium through the use of a phase diagram. Setting equation (12) equal to zero yields a schedule in the r, k plane along which $\dot{k} = 0$; this is KK in Figure 2. The slope of this schedule, obtained by differentiating $\dot{k} = 0$, is

$$(13)\quad (dr/dk)/KK = -q_2/q_1 > 0.$$

The schedule is a locus of points along which $q_N = q_S$, that is, the marginal revenue product of capital is the same in North and South. It must be upward sloping because an increase in r, *ceteris paribus*, raises the marginal revenue product of capital in the North (by improving the North's terms of trade) and this requires a flow of capital from South to North (an increase in k) to restore equilibrium.

Setting equation (11) equal to zero provides a schedule in the r, k plane along which $\dot{r} = 0$. The slope of this schedule is

$$(14)\quad \frac{dr}{dk}/RR = \frac{(1+r)f'p_2}{(i-f)-(1+r)f'p_1}.$$

[11]The assumption that $g(1) = 0$ is not crucial for the results of the model. Because of risk, capital flow from North to South may be halted when the return to capital is lower in the North than in the South, so that $g(q^*) = 0$, $q^* < 1$. In this case, all of the important results of the model still hold.

[12]The assumptions about technology diffusion and capital mobility imply that these processes are independent. The model of North-South technology transfer in Findlay (1978) links the processes in an interesting way, making the rate of technology transfer a positive function of the amount of direct foreign investment in the South.

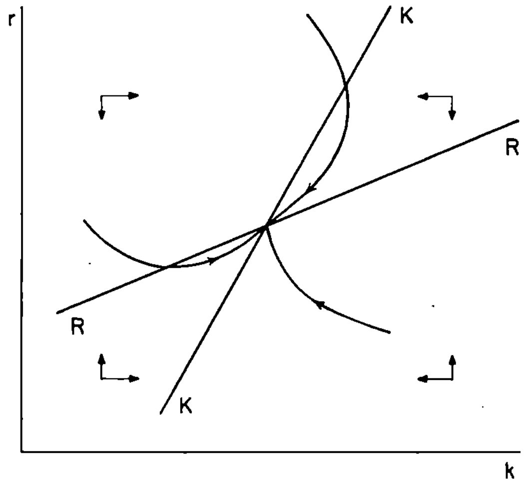


FIGURE 2

The numerator on the right-hand side of equation (14) is negative since $f'>0$ and $p_2<0$. The denominator is the partial derivative of $\dot{r}$ with respect to r, which has two components of opposite sign. From equation (11) it is clear that, holding the $f(\cdot)$ function constant, the effect on $\dot{r}$ of an increase in r is $(i-f)$, which is positive in the vicinity of an equilibrium at which $\dot{r}=0$. The negative component of the denominator in equation (14) arises because $f(\cdot)$ in fact is not constant: an increase in r raises p and hence $f(p)$, speeding up the rate of diffusion and reducing $\dot{r}$. A necessary condition for local stability of long-run equilibrium is that the latter effect dominate the former so that the partial derivative of $\dot{r}$ with respect to r is negative.[13]

In this case, the slope of the RR schedule along which $\dot{r}=0$ is positive: if from a point at which $\dot{r}=0$ r increases, $\dot{r}$ becomes negative because p rises ($p_1>0$) and this increases the rate of technology transfer to the South. An increase in k is required to reduce p ($p_2<0$), slowing down the diffusion of technology and restoring $\dot{r}=0$. Hence the RR schedule is upward sloping.

[13] Necessary and sufficient conditions for local stability of long-run equilibrium are derived in the Appendix.

Long-run equilibrium occurs at an intersection of the two schedules. The necessary and sufficient condition for local stability of such an equilibrium is that the RR schedule be upward sloping but flatter than the KK schedule in the vicinity of equilibrium, as in Figure 2. Now let us examine the adjustment path in the vicinity of a stable equilibrium. Above (below) RR, r is decreasing (increasing). To the left (right) of KK, k is increasing (decreasing). These movements are indicated by the arrows in Figure 2. Several illustrative adjustment paths are shown in the figure.

Note that while for expositional purposes it is useful to assume that capital adjusts after the variables in the static model, the *slow adjustment* of capital plays no crucial role in the analysis. The KK schedule can be interpreted as showing, for a given value of r, the corresponding division of the world capital stock that makes the marginal revenue product of capital equal in North and South. With *instantaneous* adjustment of capital the world economy would always be on the KK schedule, moving up or down it as r adjusts.

V. Characteristics of Long-Run Equilibrium

At the long-run equilibrium, both $\dot{r}$ and $\dot{k}$ are equal to zero, so that r and k have attained stable values. Thus there is a fixed division of the world's capital between North and South and a stable ratio of the number of goods produced in the North to the number produced in the South. With r and k constant, factor prices and the terms of trade also stabilize.

By assumption, $\dot{k}=0$ when $q=1$, so that at the long-run equilibrium the return to capital is equalized in the North and South; this is why the capital flow stops. Also by assumption, there is continuous introduction of new products. Thus for $\dot{r}$ to be equal to zero, there must also then still be technology transfer in the long-run equilibrium; that is, $f(p)$ is positive, and hence p is greater than 1. Intuitively, for the terms of trade and factor prices to stabilize, there must be a constant ratio of the number of goods pro-

duced in each region. Since the North is continuously introducing new goods, however, there must be a steady flow of technology to the South to stabilize r. For there to be an *incentive* for such a flow, the long-run equilibrium terms of trade must be greater than 1 so that costs of production are lower in the South than in the North for goods that can be produced in both regions.[14]

The fact that the terms of trade are greater than 1 while the returns to capital are equalized means that wages must be higher in the North than in the South; that is, for there to be a difference in costs of production the internationally immobile factor of production must earn more in the North than in the South. Thus the ability of the North to introduce new products and monopolize them temporarily enables workers in the North to earn a premium over the wages of their counterparts in the South, even though labor is assumed to be of the same quality in the two regions. In addition, the fact that w_N/q_N is greater than w_S/q_S means that less labor per unit is used in the production of a representative good in the North than in the South; goods produced in the North will be observed to be more capital intensive than goods produced in the South, though all goods in fact can be produced with the same production function.

Furthermore, the overall ratio of capital to labor will have to be higher in the North than in the South at the long-run equilibrium. It is interesting that even when capital movement is allowed to proceed to the point where the return to capital is equalized in the two regions, differences in relative factor endowments develop as an endogenous result of the model. Thus the model leads to an equilibrium at which the overall capital-labor ratio is higher in the North than in the South, and at which the North exports "capital-intensive" and imports "labor-intensive" goods, as would be expected from the factor-proportions model; but this is not the result of initial differences in factor endowments. Rather it comes about because of the North's ability to introduce new products.

As noted above, in the long-run equilibrium there will continue to be innovation and technology transfer so that the number of goods produced in each region will be constantly increasing—but at the same percentage rate so that the ratio of the number produced in each region is stable. Thus while factor prices and the terms of trade will be stable in the long-run equilibrium, the pattern of trade will be in constant flux. There will be a product cycle in which goods produced and exported by the North in one period will be produced and exported by the South in a later period. There will be stability in various macro aggregates—levels of employment, national incomes (measured in terms of the numeraire), the value of trade—while at the same time there will be constant turmoil at the micro level.

VI. Shocks to Long-Run Equilibrium

The long-run equilibrium can be disturbed by changes in the southern labor force or by an exogenous shift in the technology diffusion function, $f(p)$. In this section, I examine in detail how the values of all of the variables change as the system moves to a new long-run equilibrium with a larger labor force in the South. The effect of a shift in the diffusion function is then considered briefly.

A. *Labor Force Growth in the South*

An increase in the southern labor force leads in the long run to a decrease in r, the ratio of the number of goods produced in the North to the number produced in the South.

[14]As noted in Section I, it is possible in the short run that there would be a trading equilibrium at which the North produces both old and new goods (incomplete specialization). In this case, the consumption equilibrium is uniquely determined but the production equilibrium is indeterminate. In terms of Figure 1, this case corresponds to a demand schedule for new goods that intersects the supply schedule on the elastic portion so that $p=1$. From equation (11) it is clear that such an equilibrium cannot persist in the long run: with $p=1$, r is increasing, which shifts the demand schedule for new goods to the right continuously. This shifting will only cease if $\dot{r}=0$, which implies that $p>1$. Thus in the long-run equilibrium there must be complete specialization.

In moving from one long-run equilibrium to another, the change in r in response to a change in L_S is

$$(15)\quad dr/dL_S = \frac{(1+r)f'(q_3p_2 - q_2p_3)}{(1+r)f'(p_1q_2 - p_2q_1) - (i-f)q_2} < 0.$$

If the initial long-run equilibrium is locally stable, the denominator of this fraction is negative. The numerator is positive because of the relationships among the partial derivatives of p and q established in Section II.

An increase in the southern labor force also leads in the long run to a decrease in k, the ratio of the stock of capital in the North to the stock of capital in the South. The change in k in response to a change in L_S as the system moves from one long-run equilibrium to another is

$$(16)\quad dk/dL_S = \frac{[(i-f)-(1+r)f'p_1]q_3 + (1+r)f'q_1p_3}{(1+r)f'(p_1q_2 - p_2q_1) - (i-f)q_2} < 0$$

Again the stability condition ensures that the denominator is negative. If $q_3 < 0$, then both terms in the numerator are positive. If $q_3 > 0$, the numerator is still positive; to see this rearrange it as $(i-f)q_3 + (1+r)f'(q_1p_3 - p_1q_3)$. The first term is positive since $(i-f) > 0$ and the second term is positive as well because $q_1p_3 > p_1q_3$.

We can use Figure 3 to examine the effect of an increase in the Southern labor force. An increase in L_S shifts the RR schedule down to $R'R'$. The KK schedule will shift left or right depending on whether q_3 is negative or positive; but, if it shifts to the right, the shift must be small relative to the shift in the RR schedule. Either way the new equilibrium occurs at a point at which both r and k are smaller.

In Figure 3 it is assumed that $q_3 < 0$ so that KK shifts to the left to $K'K'$. The original equilibrium point is now in the region where both r and k are declining. This is because the impact effect of the increase in L_S is to increase p and reduce q. This in-

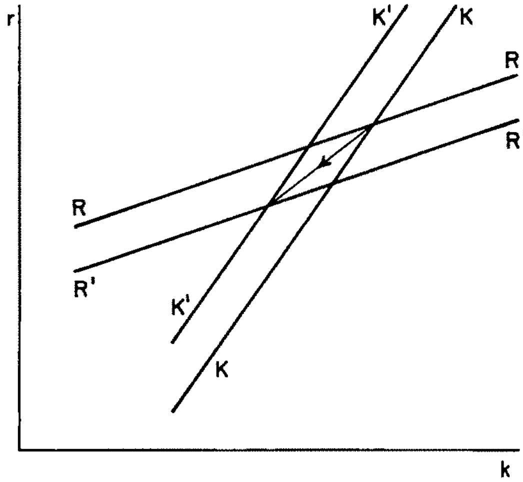


FIGURE 3

creases the flow of technology, so that r begins to decline. Also, the decrease in q initiates a movement of capital from the North to the South, decreasing k. Thus over time the increase in the southern labor force leads to a flow of capital and an increased flow of technology from the North to the South.

The new schedules, $K'K'$ and $R'R'$, intersect at a point at which both r and k have smaller values than at the initial equilibrium. In moving to the new long-run equilibrium, q must rise back to 1, since this is the only value at which $\dot{k} = 0$. Turning to the $\dot{r} = 0$ equilibrium condition, this equation can be expressed as

$$(17)\quad \dot{r} = [i - f(p)]r - f(p) = 0.$$

In moving from one equilibrium to the other, i does not change and r declines; thus $f(p)$ must decline for this equation to be satisfied. This implies that the terms of trade, p, are lower at the new equilibrium than at the old.

The impact effect of an increase in the southern labor force is to improve the North's terms of trade and to reduce the return to capital in the North relative to the return in the South (assuming $q_3 < 0$). These changes, however, initiate a flow of capital and an increased flow of technology; over time these

flows lead to an equalization of the return to capital in the two regions and to a decline in the terms of trade to a value lower than its value at the initial equilibrium.

Let us turn now to the changes in factor prices that result from an increase in the southern labor force. An increase in L_S initially increases the northern wage by improving the North's terms of trade. As r and k adjust downward, however, the capital-labor ratio in the North falls, reducing w_N/p, the product wage (marginal physical product of labor) in the North. Since p declines in moving to the new long-run equilibrium, the long-run change in w_N also must be negative. Thus an increase in the southern labor force leads eventually to a decrease in northern wages relative to both types of consumption goods.

This result is somewhat counterintuitive given the classical nature of the model. As noted, the impact effect of an increase in the labor force in the South is to raise wages in the North, as is generally the case in classical models with complete specialization. In this model, however, the short-run changes in factor prices and the terms of trade that result from an increase in L_S lead in the long run to a flow of capital and technology from the North to the South, eventually undermining northern wages. The mobility of capital and technology creates a *tendency* toward equalization of wages in the two regions (though as long as there is innovation in the North, equalization of wages will never actually be realized). Thus over time the increased supply of labor in the South puts downward pressure on wages in the North, as is the case in the neoclassical trade model.[15]

It should be emphasized that the long-run decline in wages in the North is the result of two changes, a reduction in Northern labor's marginal physical productivity and a decline in the North's terms of trade. The long-run effect on the North's terms of trade of an increase in L_S depends on the specific form of the assumptions about innovation and diffusion. Altering the assumption in equation (8), for instance, to $\dot{n} = in$ would make long-run equilibrium p independent of L_S. In this situation an increase in L_S still reduces northern wages in the long run by inducing an outflow of capital from the North and thus reducing northern labor's marginal physical productivity. Hence the inclusion in the model of capital as a second factor of production that is internationally mobile is crucial for establishing the robustness of the negative relationship between northern wages and the supply of labor in the South.

The long-run changes in other factor prices resulting from an increase in L_S cannot be established as definitively as the change in w_N. The impact effect of an increase in L_S is to increase q_S and reduce w_S; over time, however, as capital flows into the South, q_S declines and w_S increases. The changes in q_S and w_S in moving from one long-run equilibrium to another are hence ambiguous.

B. *An Exogenous Increase in the Rate of Diffusion*

Long-run equilibrium will also be disturbed if there is an exogenous increase in the rate of diffusion, increasing the $f(p)$ function for any value of p. This will shift the RR schedule down so that both r and k decline in moving to the new long-run equilibrium.

The North's terms of trade decline and the North's capital-labor ratio falls; thus more rapid diffusion has the same effect on northern wages as labor force growth in the South.

The effect on other factor prices, however, is different. The induced flow of capital to the South now unambiguously increases the capital-labor ratio in the South. Thus more rapid diffusion of technology raises real wages in the South for two reasons: it leads to improvements in the South's terms of trade and in the marginal physical productivity of southern workers.

[15] The mechanism through which this occurs in the neoclassical model is of course different. The increased labor force can lead to a reduction in wages in the country (or region) where it occurs. If factor-price equalization obtains, there will have to be a reduction in wages in the other country (or region) as well. This result can be derived from the general equilibrium model presented in Jones (1965).

VII. Conclusions

The model constructed in this paper, like other models in the pure theory of international trade, is highly stylized and in some respects unrealistic. Nevertheless, it captures some of the forces that are shaping the pattern of trade in the real world today—forces that cannot be easily accommodated within the framework of the neoclassical factor-endowment theory of international trade. As such, the model is intended as a supplement to, not a replacement for, that conventional theory.

The model depicts a world in which the pattern of trade is in constant flux. New products and whole industries develop in the industrial countries while other industries there decline in the face of low-cost competition from the developing nations. By tying the *pace* of innovation, diffusion, and capital movement to the prices determined through trade, the model implies that over time the technology and capital that nations have available will be as much the result as the cause of trade. In this way the model takes a step toward making structural change in the world economy (and hence in the pattern of trade) an endogenous element of trade theory.

To note that in the developed countries, industries like steel, automobiles, and textiles are losing out to Third World competitors and that differences in wages play a key role in this process is commonplace in the business press. The problem with these analyses is that they tend to take the wage gap as given, when of course international wage comparisons can only be made through the exchange rate, which is the endogenous result of trade in goods (and other international transactions). The model constructed in this paper pinpoints innovation in the North as the factor that enables and in fact *requires* that workers in the North earn a premium over wages in the South, even though the quality of labor is assumed to be the same in the two regions.

The main insight of the model is that for factor prices and the terms of trade to be stable, there must be a stable ratio of the number of goods produced in each region. If the North is constantly introducing new products, however, the ratio can be stable only if there is simultaneously a flow of technology to the South. For there to be an incentive for such a flow, labor, the internationally immobile factor of production, must earn a greater reward in the North than in the South. Thus constant innovation is necessary just to *maintain* living standards in the North; in the absence of innovation the transfer of capital and technology to the South would undermine northern wages.

Note that it is the cost-driven technology transfer (combined with constant innovation in the North) that really *necessitates* wages being higher in the North than in the South in long-run equilibrium. Paradoxically, it is this same mechanism that undermines northern wages when the labor force expands in the South. Labor force growth in the South initially leads to higher wages in the North, but over time this attracts a flow of capital and an increased flow of technology from the North that reduces wages there.

Thus international mobility of capital and technology creates a *tendency* toward equalization of all factor prices, though as long as there is innovation in the North, strict equalization of wages will never occur. In this way the model preserves a number of attractive features of the classical model especially relevant to trade between developed and less developed countries: differences in technology play a crucial role in determining the pattern of trade, the two regions in equilibrium produce different sets of goods, and wages in the two regions can be different (and heavily dependent on the terms of trade). At the same time the model illustrates that trade in goods combined with some international factor mobility creates a tendency toward factor-price equalization analogous to the neoclassical result.

That labor force growth in the less developed countries (combined perhaps with exogenous increases in the rate of diffusion) must eventually reduce demand for labor in the industrial countries by accelerating the transfer of technology appears consistent with recent events. Much of the production in industries like steel, textiles, automobiles, and electronics has moved to the South; new

industries, however, have not arisen at a fast enough pace to employ all of the displaced workers in the developed countries. This process has no doubt been a prime cause of the high levels of unemployment in the industrial nations over the last decade.

Real wages have declined slowly in the face of the high unemployment. Resistance to wage cuts has often focused on the role that imports from the less developed countries play in reducing demand for industrial labor in the developed countries. The result of this protest has been more and more restrictions on trade in goods, usually in the form of quotas.

The continued rapid growth of the labor force in the South—combined with capital and technology transfer and trade in goods—should put further downward pressure on wages in the North, leading to even more demands for protection from imports. The main policy implication of the model would then seem to be that some government control of the process of capital and technology transfer may be desirable in order to prevent further erosion of the world's relatively open trade in goods.

APPENDIX

Stability of Long-Run Equilibrium

The long-run equilibrium of the system is defined by the two equations

$$\text{(A1)} \quad \dot{r} = ir - (1+r)f(p(r,k,L_S)) = 0,$$

$$\text{(A2)} \quad \dot{k} = g(q(r,k,L_S)) = 0.$$

The characteristic equation corresponding to this system is

$$\text{(A3)} \quad \lambda^2 - [(i-f)-(1+r)f'p_1 + g'q_2]\lambda - g'[(1+r)f'(p_1q_2 - p_2q_1) - (i-f)q_2] = 0.$$

A necessary and sufficient condition for the local stability of equilibrium is that all coefficients of this equation be positive. Since $g' > 0$, the constant term will be positive if and only if $[(1+r)f'(p_1q_2 - p_2q_1) - (i-f)q_2]$ is negative. This expression can be rewritten as $[(1+r)f'p_1 - (i-f)]q_2 - (1+r)f'p_2q_1$; since $q_1 > 0$, $q_2 < 0$, and $p_2 < 0$, $[(1+r)f'p_1 - (i-f)]$ must be positive if the whole expression is negative, that is, if the equilibrium is stable.

The requirement that the constant term in (A3) be positive also implies that

$$\text{(A4)} \quad \frac{(1+r)f'p_2}{(i-f)-(1+r)f'p_1} < \frac{-q_2}{q_1},$$

where both of these expressions are positive. The left-hand side of inequality (A4) is the slope of the *RR* schedule; the right-hand side is the slope of the *KK* schedule. Thus stability requires that *KK* be steeper than *RR* in the vicinity of equilibrium.

REFERENCES

Dixit, Avinash and Stiglitz, Joseph E., "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Dollar, David, "Essays in the Pure Theory of International Trade: The Effects of Economies of Scale and Technological Innovation on the Pattern of Trade," unpublished doctoral dissertation, New York University, May 1984.

Dornbusch, Rudiger, Fischer, Stanley and Samuelson, Paul, "Comparative Advantage, Trade, and Payments in a Ricardian Model with a Continuum of Goods," *American Economic Review*, December 1977, *67*, 823–39.

Findlay, Ronald, "Relative Backwardness, Direct Foreign Investment, and the Transfer of Technology: A Simple Dynamic Model," *Quarterly Journal of Economics*, February 1978, *92*, 1–16.

———, "The Terms of Trade and Equilibrium Growth in the World Economy," *American Economic Review*, June 1980, *70*, 291–99.

Krugman, Paul, "A Model of Innovation, Technology Transfer, and the World Distribution of Income," *Journal of Political Economy*, April 1979, *87*, 253–66.

______, "Trade, Accumulation, and Uneven Development," *Journal of Development Economics*, April 1981, *8*, 149–61.

Jones, Ronald, "The Role of Technology in the Theory of International Trade," in R. Vernon, ed., *The Technology Factor in International Trade*, Universities–National Bureau Conference Series, No. 22, Columbia University Press, 1970.

______, "The Structure of Simple General Equilibrium Models," *Journal of Political Economy*, December 1965, *73*, 557–72.

Vernon, Raymond, "International Investment and International Trade in the Product Cycle," *Quarterly Journal of Economics*, May 1966, *80*, 190–207.

Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data

By RATI RAM*

A study of the impact of government size on economic performance and growth is important. Theoretically, one point of view suggests that a larger government size is likely to be detrimental to efficiency and economic growth because, for example, (*i*) government operations are often conducted inefficiently, (*ii*) the regulatory process imposes excessive burdens and costs on the economic system, and (*iii*) many of government's fiscal and monetary policies tend to distort economic incentives and lower the productivity of the system. At the other extreme, one can identify some points of view that assign to the government a critical role in the process of economic development, and could argue that a larger government size is likely to be a more powerful engine of economic development. There are several arguments on which the latter point of view is based. These include, besides others, (*i*) role of the government in harmonizing conflicts between private and social interests, (*ii*) prevention of exploitation of the country by foreigners, and (*iii*) securing an increase in productive investment and providing a socially optimal direction for growth and development.

Although the subject is important and theoretical positions can be quite diverse, direct empirical assessments of the issue seem to have been few, and two fairly recent studies reported contradictory results. Using a sizable cross-country sample, Richard Rubinson (1977) concluded that a larger government size, indexed by the share of government revenue in *GNP*, promotes economic growth by reducing "dependence," especially in the poorer, less developed contexts. On the other hand, Daniel Landau (1983) concluded that a larger government size, proxied by the share of government consumption in *GDP*, depresses growth of per capita income.[1]

Several features characterize this study. The specifications used are derived from production function modeling of government and nongovernment sectors, and the estimated growth models not only provide an assessment of the overall effect of government size on economic growth, but also enable one to judge whether (a) the (marginal) "externality" effect of government size on the rest of the economy is positive or negative, and (b) input productivity in the government sector is higher or lower than in the nongovernment (private) sector. Specification tests for nonnested models proposed by Russell Davidson and James MacKinnon (1981) are used to assess the goodness of the main models used in the study, relative to a widely used conventional specification. Internationally comparable data on output, investment, and government size, published by Robert Summers and Alan Heston (1984) for a large sample of 115 countries, are used, and a full two-decade period from 1960 through 1980 is covered. In addition to the cross-section analysis, from which almost all previous studies derive their results, a beginning is made toward obtaining estimates based on time-series data for a sizable number of individual countries.

The main results are: 1) the overall impact of government size on growth is positive in almost all cases; 2) the (marginal) externality

*Department of Economics, Illinois State University, Normal, IL 61761. T. W. Schultz gave insightful comments on an earlier version. Rosa Lea Danielson and Ann Sullivan provided competent research assistance. Partial support from Graduate School of Illinois State University is gratefully acknowledged. I alone am responsible for any errors or deficiencies.

[1] There are, of course, other studies that bear directly or indirectly on the issue. Some of the studies mentioned by P. C. Afxentiou (1982) are related to the question addressed in this work.

effect of government size is generally positive; 3) compared with the rest of the economy, factor productivity in the government sector appears to be higher, at least during the 1960's; 4) although the number of time-series observations for each country is relatively small, there is a broad harmony between the estimates obtained from cross-section and time-series data; 5) as compared with the 1960's, the positive externality effect of government size on growth may have become somewhat stronger during the 1970's, but relative factor productivity in government sector may have declined over the 1970's; 6) the Davidson-MacKinnon tests indicate the models used in the study are preferable to a widely used alternative; and 7) it is possible that the positive effect of government size on growth is stronger in lower-income contexts.

Section I outlines the theoretical framework and the models used. Section II describes the data and the variables. Section III reports and discusses the main cross-section results. Section IV deals with the model specification tests. Section V reviews briefly the estimates based on time-series data.

I. The Theoretical Framework

The two-sector production function framework outlined in this section is adapted from the reasoning developed by Gershon Feder (1983, pp. 61–67), and seems to provide an appealing set of models for investigating the relation between government size and growth of aggregate output.

Assume the economy consists of two broad sectors, the government sector (G) and the nongovernment sector (C). If output in each sector depends on the inputs of labor (L) and capital (K), and if, in addition, output ("size") of the government sector exercises an "externality" effect on output in the other sector (C), production functions for the two sectors may be written as

$$C = C(L_c, K_c, G), \tag{1}$$

$$G = G(L_g, K_g), \tag{2}$$

where subscripts denote sectoral inputs. If the total inputs are given,

$$L_c + L_g = L, \tag{3a}$$

$$K_c + K_g = K. \tag{3b}$$

The total output (Y) is just the sum of outputs in the two sectors, and thus

$$C + G = Y. \tag{3c}$$

Let the relative factor productivity in the two sectors differ; in particular,

$$G_L/C_L = G_K/C_K = (1+\delta), \tag{4}$$

where uppercase subscripts denote partial derivatives of the functions with respect to subscripted input; for example, G_L denotes $\partial G/\partial L$ or its discrete analog $\Delta G/\Delta L$. It is obvious that the sign of δ indicates which sector has higher marginal factor productivity, and a positive δ implies higher input productivity in the government sector. By manipulating the production functions, and using (3) and (4), the following approximation for an aggregate growth equation can be derived:

$$\dot{Y} = \alpha(I/Y) + \beta\dot{L} + [(\delta/(1+\delta)) - \theta]\dot{G}(G/Y) + \theta\dot{G}, \tag{5}$$

or, writing δ' for $\delta/(1+\delta)$,

$$\dot{Y} = \alpha(I/Y) + \beta\dot{L} + (\delta' - \theta)\dot{G}(G/Y) + \theta\dot{G}, \tag{5'}$$

where a dot over the variable indicates its rate of growth; for example, $\dot{Y}$ denotes dY/Y or its discrete equivalent $\Delta Y/Y$. The parameters β, α, and θ are of the kind usually found in simple aggregate growth models. In this case, β is the elasticity of nongovernment output C with respect to L; α is the marginal product of K in the C sector; and θ equals $C_G(G/C)$, and is the elasticity of nongovernment output with respect to G.[2]

[2] More details about the derivations, and the interpretation of the models and the parameters, are in Feder (pp. 61–67).

The variable I is investment and, as usual, is assumed to equal dK (or ΔK). If θ is believed to be a constant parameter across the sample observations, equation (5) provides an econometric specification that can easily yield estimates of δ and θ which indicate, respectively, the intersectoral factor productivity difference and the marginal externality effect of government output (i.e., size) on the rest of the economy and hence on economic performance. It may be noted that both C_G and θ represent the externality affect; C_G is similar to "marginal product," and gives, for constant L_c and K_c, the increase in non-government output (and hence the total output) as G, the government size, increases by one unit. The parameter θ is an elasticity measure, and reflects the percentage increase in C (for given L_c and K_c) with a 1 percent increase in G.[3]

One special case of (5) is noteworthy. If $\delta' = \theta$, (5) reduces to[4]

$$(6) \qquad \dot{Y} = \alpha(I/Y) + \beta\dot{L} + \theta\dot{G}.$$

In (6), as in (5), θ gives only the externality effect of government size, and not the total effect. However, since (6) is premised on $\delta' = \theta$, estimate of θ also yields an estimate of δ' (and of δ), and therefore of the total effect, provided the constraint $\delta' = \theta$ is valid.

If one prefers to postulate that C_G, rather than θ, is the constant parameter, (5) can be rewritten as[5]

$$(7) \quad \dot{Y} = \alpha(I/Y) + \beta\dot{L} + (\delta' + C_G)\dot{G}(G/Y).$$

Clearly, the coefficient of $\dot{G}(G/Y)$ in (7) is very different from the coefficient of that variable in (5); typically, the coefficient in (5) is likely to be much smaller than in (7). Also, it is obvious that the advantage of estimating (7) is that, unlike (5) or (6), one can obtain the overall effect of government size directly from the coefficient of $\dot{G}(G/Y)$; but the disadvantage is that one cannot get separate estimates of the externality effect and the factor productivity differential. Another point to note is that while collinearity between $\dot{G}$ and $\dot{G}(G/Y)$ may lower precision in the estimation of (5), neither (6) nor (7) has that drawback. Comparing the last two, although simple and informative, (6) is premised on the equality of δ' and θ; on the other hand, although less informative in regard to separate estimates of δ and C_G, (7) does not rest on parametric constraints of the type (6) involves.

The main focus of this study is on obtaining at least direction of the overall effect of government size on growth, and sign of the marginal externality effect parameter (θ or C_G) and of the intersectoral productivity differential (δ). The broad strategy, there-

[3]Obviously, therefore, both parameters indicate the marginal externality effect of government size, and do not merely reflect the externality effect of the "existence" of a government.

[4]If $\delta = 0$ but $\delta' \neq \theta$, (5) can be written as

$$\dot{Y} = \alpha(I/Y) + \beta\dot{L} + \theta[\dot{G} - \dot{G}(G/Y)],$$

or $$\dot{Y} = \alpha(I/Y) + \beta\dot{L} + \theta\dot{G}(C/Y).$$

In this equation, but not in (6), θ indicates the total effect of government size since $\delta = 0$. It is obvious, of course, that if $\delta = 0$, *and* $\delta' = \theta$, both δ and θ (and C_G) would be zero, and government size would have no effect on growth in this framework.

[5]Specifications that formally resemble (6) and (7) can be directly derived through the somewhat traditional process of introducing G as an "input" in the aggregate production function $Y = f(L, K, G)$, which, on taking total derivatives and manipulation of the expression, leads to the "standard" form:

$$\dot{Y} = \alpha_K(I/Y) + \beta_L\dot{L} + \beta_G\dot{G},$$

or equivalently,

$$\dot{Y} = \alpha_K(I/Y) + \beta_L\dot{L} + \alpha_G(\Delta G/Y).$$

The parameters in the above equations have slightly different meanings from those in (6) and (7), but the formal similarity becomes obvious on noting that $\dot{G}(G/Y)$ equals $\Delta G/Y$. The main appeal of the models proposed in the text lies in the information they can convey about the mechanisms through which government size may affect economic growth; in particular, (*i*) the externality effect of government size is explicitly modeled, and there is no *ad hoc* introduction of G as an input in the aggregate output equation, and (*ii*) intersectoral productivity differential is explicitly allowed for, and can be estimated. If $\delta = 0$ in (5), there is really not much difference between the specifications in the text and the ones given above; but the former still have the merit of explicitly modeling and estimating the externality effect.

fore, is to consider the estimated coefficient of $\dot{G}(G/Y)$ in (7) to assess directly the overall effect. Also, since the estimation of (5) suggests that $\delta' - \theta = 0$, (6) is equivalent to (5), θ (and C_G) and δ have the same sign, and an indication of both is obtained from the coefficient of $\dot{G}$ in (6).[6] Of course, if the restriction $\delta' = \theta$ is not valid, and one has a sample that is large enough to permit reasonably precise estimation of (5), it is possible to get signs (and magnitudes) of δ and θ from (5).

It should be noted that in none of the specifications developed in this section, nor in those based on homogeneous aggregate production functions that include G as an input, does the ratio G/Y appear as an independent variable by itself; one gets either $\dot{G}$ or $\Delta G/Y$ (or dG/Y) which is the same as $\dot{G}(G/Y)$.[7] However, as the studies by Rubinson and Landau indicate, specifications that include a regressor like G/Y seem to be widely used for assessing the impact of government size on economic growth or development. Therefore, some estimates for the growth model

$$(8) \qquad \dot{Y} = \alpha_K(I/Y) + \beta_L\dot{L} + \gamma(G/Y)$$

are also reported so as to compare these with the results obtained from (6) and (7).

II. Data and the Variables

Since the investigation involves cross-country comparisons, international comparability of the data is of obvious importance, especially when the focus is on a cross-section analysis. Typically, compared with more developed economies, domestic prices of government services are considerably lower, and of investment goods higher, in less developed countries (*LDC*s). The report by Irving Kravis et al. (1982, especially p. 21) shows dramatically the large cross-country variations in the relative prices of investment goods and government services. Internationally comparable data on output, investment, and government services (sometimes called "government consumption") have been rare. However, Summers and Heston have recently published a valuable data set that contains annual estimates of output, investment, consumption, and government services on an internationally comparable basis. They have updated to 1980 the data published earlier in Summers et al. (1980) and have made improvements and corrections in the earlier estimates. The Summers-Heston data are used in this study.

Rate of increase of *GDP* is taken as a proxy for economic growth, and *GDP* in "international dollars" is used for the aggregate output measure Y. The Summers-Heston data on gross domestic investment and on "government" are used for I and G, respectively, and both seem to be reasonably good proxies for the theoretical variables. As in several other studies, rate of population growth ($\dot{P}$) is used in place of the rate of increase in labor input ($\dot{L}$). Although not really a good proxy in some cases, use of $\dot{P}$ does have some advantages. Good time-series data on labor force are rare, particularly for the *LDC*s, but data on population are fairly good; with $\dot{P}$ as a regressor in the growth model, the specification resembles one in which growth of per capita output is the dependent variable, since the latter can be considered as a special case in which the coefficient of $\dot{P}$ is constrained to unity; and availability of information on population in the Summers-Heston data set makes for uniformity of the source for all variables.

The Summers-Heston data set includes 115 market economies. While a large number of entries are missing for years prior to 1960, complete data are available from 1960 through 1980, which is the period studied in this work.

Cross-section estimates are based on mean values of I/Y and G/Y for the periods considered, and rates of growth of Y, P, and G are computed by fitting exponential trend equations to variable values for the period. To bring out any possible structural variations between the 1960's and the 1970's, sep-

[6]Ruling out negative marginal factor products in the government sector, and therefore assuming $\delta \nless -1$, it is easily seen that δ' and δ would have the same sign and would be monotonically related.

[7]See also the equations given in fnn. 4 and 5.

arate estimates are reported for 1960–70 and 1970–80.

Estimates based on time-series data cover the full 1960–80 period, so that there are 20 observations for each country. Annual rates of growth are approximated by first differences of the logarithms of the variable values for successive years.

III. The Main Cross-Section Results

Table 1 contains the principal results from cross-section data. The estimates are given for equations (5), (6), (7), and (8), separately for 1960–70 and 1970–80, and four different country groups are included. A constant term is added, and a random stochastic disturbance term with the usual nice properties is assumed. Before the results are discussed, it might be useful to review the expected pattern of the coefficients of $\dot{G}$ and $\dot{G}(G/Y)$ in (5), (6), and (7). As one looks at the equations, it should be obvious that for positive values of θ, the coefficient of $\dot{G}(G/Y)$ in (7) is likely to be much larger than the coefficient of that variable in (5); also the former is expected to be considerably larger than the coefficient of $\dot{G}$ in (6) unless δ is a substantial negative number.[8] Illustratively, if $\delta = \theta = 0.25$ and $G/Y = 0.20$, the coefficients of $\dot{G}(G/Y)$ and $\dot{G}$ in (5) would be -0.05 and 0.25, respectively, that of $\dot{G}(G/Y)$ in (7) would be 1.20, and that of $\dot{G}$ in (6) may be somewhere around 0.25.

Subject to the usual caveats appropriate to such cross-section studies, the following points emerge from the estimates reported in Table 1.[9]

The coefficient of $\dot{G}(G/Y)$ in equation (5) is not significantly different from zero in any of the eight cases at any reasonable significance level. Therefore, one can drop that term and focus on the simpler version (6). Also, since it follows that the constraint $\delta' = \theta$ is valid, δ and θ are likely to have the same sign, and one can use (6) for obtaining an indication of the direction of the externality effect of government size on growth and also of the intersectoral factor productivity differential.

Estimates of the coefficient of $\dot{G}$ in (6) are positive and statistically significant at least at the 1 percent level in every case. Therefore, it is fair to conclude that the externality effect of the government size on the rest of the economy, and hence on economic growth, is positive in all cases for both periods.[10]

Since δ and θ are likely to have the same sign, it follows that δ is also positive, and factor productivity in government sector is higher than in the nongovernment (private) sector. For 1960–70, that conclusion is justified not only on the basis of the sign of θ in (6) and the implied equality of δ' and θ, but also if one takes seriously the nonsignificant estimates of $(\delta' - \theta)$ in equation (5). For 1970–80, while the positive sign of θ, along with the equality of δ' and θ, implies a positive sign on δ, the nonsignificant point estimates of the coefficient of $\dot{G}(G/Y)$ in (5) seem to carry a different implication. In other words, although a proper mode of statistical inference suggests δ is probably positive for both periods, one can adopt a more cautious approach, and conclude that input productivity in government sector was higher than in the rest of the economy at least during the 1960's.[11]

[8] See fn. 6. Even when $\delta \nless -1$, δ' can take large negative values as δ becomes smaller than -0.5; for instance, when $\delta = -0.8$, $\delta' = -4$.

[9] Besides several other well-known reasons for caution in interpreting such results, it seems best to regard the parameter estimates as reflecting intercountry averages, and not to take the cross-section models too strictly in the sense of treating the estimates as applicable to each sample country. For considerations of this nature, the focus of this study is on parameter signs and not on the estimated numbers.

[10] One should perhaps not conclude that increase in every government activity is beneficial for the rest of the economy, but only that the net externality effect is positive. If expansion of some activities causes harm, their effect is dominated by others that help the nongovernment sector.

[11] Computational details that enable one to relate the estimates from (5), (6), and (7) for getting values of δ and θ are quite simple, and are omitted to save space. The basic approach is to adopt the more conservative conclusion. Also, if one uses the point estimates in (5), (6), and (7), it is possible to make a variety of calculations regarding the productivities of labor and capital, the externality effect parameters (C_G and θ), and the

TABLE 1—ESTIMATES OF EQUATIONS (5) THROUGH (8) FOR 1960–70 AND 1970–80 FOR FOUR COUNTRY GROUPS[a]

	1960–70						1970–80					
Equations	I/Y[b]	$\dot{P}$[c]	$\dot{G}(G/Y)$[d]	$\dot{G}$[e]	G/Y[f]	$R^2(F)$[g]	I/Y[b]	$\dot{P}$[c]	$\dot{G}(G/Y)$[d]	$\dot{G}$[e]	G/Y[f]	$R^2(F)$[g]
I. Entire Sample												
All *LDC*s and *DC*s ($N=115$)												
(5)	0.114[h]	0.504[h]	0.672	0.139[i]		0.35	0.097[h]	0.453[i]	−0.554	0.485[h]		0.46
	(4.81)	(2.45)	(1.59)	(1.92)		(14.92)	(3.30)	(1.87)	(−0.76)	(3.41)		(23.23)
(6)	0.114[h]	0.551[h]		0.226[h]		0.34	0.100[h]	0.410		0.384[h]		0.46
	(4.79)	(2.69)		(4.77)		(18.79)	(3.44)	(1.75)		(7.29)		(30.89)
(7)	0.118[h]	0.517[h]	1.286[h]			0.33	0.119[h]	0.389	1.744[h]			0.40
	(4.96)	(2.49)	(4.63)			(18.23)	(3.98)	(1.54)	(6.19)			(24.72)
(8)	0.129[h]	0.781[h]			−0.020	0.20	0.143[h]	1.275[h]			−0.108[h]	0.23
	(4.97)	(3.54)			(−0.62)	(9.45)	(4.24)	(4.62)			(−2.13)	(10.78)
All *LDC*s ($N=94$)												
(5)	0.125[h]	0.431	0.622	0.138[i]		0.37	0.124[h]	−0.014	−0.722	0.503[h]		0.47
	(4.35)	(1.75)	(1.35)	(1.78)		(13.19)	(3.70)	(−0.04)	(−0.91)	(3.24)		(19.68)
(6)	0.131[h]	0.452[i]		0.217[h]		0.36	0.126[h]	−0.071		0.372[h]		0.46
	(4.54)	(1.83)		(4.31)		(16.82)	(3.78)	(−0.22)		(6.57)		(26.01)
(7)	0.126[h]	0.447[i]	1.246[h]			0.35	0.143[h]	−0.101	1.684[h]			0.41
	(4.33)	(1.79)	(4.13)			(16.14)	(4.11)	(−0.30)	(5.49)			(20.57)
(8)	0.155[h]	0.597[h]			−0.037	0.24	0.184[h]	0.694[i]			−0.195[h]	0.30
	(5.02)	(2.23)			(−1.02)	(9.25)	(5.01)	(1.93)			(−3.36)	(12.62)
II. Sample Limited to Countries Included in Table 4												
*LDC*s and *DC*s ($N=70$)												
(5)	0.104[h]	0.632[h]	0.640	0.172[h]		0.44	0.085[h]	0.551[h]	−1.274	0.647[h]		0.47
	(3.56)	(2.47)	(1.33)	(2.15)		(12.69)	(2.44)	(2.01)	(−1.10)	(2.85)		(14.42)
(6)	0.108[h]	0.689[h]		0.250[h]		0.42	0.092[h]	0.497[i]		0.408[h]		0.46
	(3.66)	(2.72)		(4.53)		(16.14)	(2.69)	(1.84)		(6.12)		(18.77)
(7)	0.111[h]	0.650[h]	1.392[h]			0.40	0.106[h]	0.484	1.881[h]			0.40
	(3.72)	(2.47)	(4.12)			(14.58)	(2.94)	(1.68)	(5.26)			(14.91)
(8)	0.140[h]	0.974[h]			−0.021	0.25	0.101[h]	1.261[h]			0.154[h]	0.23
	(4.27)	(3.43)			(−0.53)	(7.21)	(2.47)	(3.87)			(−2.60)	(6.67)
*LDC*s Only ($N=57$)												
(5)	0.127[h]	0.615[i]	0.438	0.187[h]		0.45	0.120[h]	−0.011	−1.740	0.734[h]		0.50
	(3.25)	(1.99)	(0.79)	(2.11)		(10.83)	(3.07)	(−0.03)	(−1.37)	(2.95)		(12.85)
(6)	0.135[h]	0.639[h]		0.237[h]		0.45	0.126[h]	−0.077		0.408[h]		0.48
	(3.57)	(2.08)		(3.85)		(14.34)	(3.22)	(−0.20)		(5.79)		(16.25)
(7)	0.130[h]	0.623[i]	1.278[h]			0.41	0.136[h]	−0.074	1.865[h]			0.41
	(3.23)	(1.95)	(3.20)			(12.16)	(3.27)	(−0.18)	(4.88)			(12.42)
(8)	0.187[h]	0.805[h]			−0.057	0.32	0.158[h]	0.602			−0.253[h]	0.33
	(4.69)	(2.37)			(−1.30)	(8.14)	(3.58)	(1.38)			(−3.80)	(8.75)

[a]A constant term *is* included in all regressions, but its estimates are not reported just to save space. The classification of countries between *LDC*s and *DC*s is based on World Bank (1984, pp. *xxiii–xxxv*). All industrial market economies are *DC*s, and developing countries are *LDC*s.

[b]The variable is investment-*GDP* ratio, and the numbers here, and all other variables, are the estimated coefficients, with *t*-statistics shown in parentheses below.

[c]The variable is the rate of population growth (a proxy for the rate of growth of labor force).

[d,e,f]The variable $\dot{G}$ is the rate of growth of government, and G/Y is the ratio of (value of) government services to *GDP*.

[g]The numbers are R^2s, with the regression *F*-statistics shown below in parentheses.

[h]The estimated coefficient is statistically significant at least at the 5 percent level.

[i]The estimated coefficient is statistically significant at the 10 percent level.

Estimates for the coefficient of $\dot{G}(G/Y)$ in equation (7), which gives the overall effect of government size on growth, are large positive numbers in every case, and are statistically significant at least at the 1 percent level. Therefore, irrespective of the position one takes in regard to the sectoral productivity differential for 1970–80, the estimates of (7) lead to the conclusion that the total effect of government size on growth is positive, and probably quite large, in every case for both periods.

intersectoral productivity differential. The approach adopted in this work is to consider the parameter signs and the broad magnitudes, and not to take the estimated parameters as exact numbers.

The coefficients of $\dot{G}$ are higher for 1970–80 than for 1960–70 in each of the sixteen cases. Therefore, although a rigorous statistical test is not offered, it seems likely that the positive externality effect of government size became stronger over the 1970's. A full investigation of, or explanation for, that possibility is not pursued here, but an increase in the externality effect could have occurred because of the important role governments could play in responding to several external "shocks" of the 1970's; for example, (a) the tremendous increase in oil prices, (b) international transmission of inflation and strains on international trade, and (c) the general balance of payment problems, especially in the *LDC*s.

If one believes in the equality of δ' and θ, implied by nonsignificance of the $\dot{G}(G/Y)$ term in (5), a higher value of θ implies a larger value for δ, which in turn suggests that relative factor productivity in government sector became higher over the 1970's, along with the positive externality effect having become stronger. However, for reasons given in the discussion relating to the sign of δ, if one takes seriously the nonsignificant estimates of the coefficient of $\dot{G}(G/Y)$ in (5) for 1970–80, it might seem that δ was smaller in 1970–80 than in 1960–70. Therefore, adopting a cautious inferential approach here also, it may be concluded that while the positive externality effect of government size may have become stronger during the 1970's, relative factor productivity in government sector could have declined over that period.[12]

The estimated coefficient of G/Y in equation (8) is negative in every case, and, while not statistically significant in the regressions for 1960–70, is significant at least at the 5 percent level in all the four cases for 1970–80. Despite several differences between the regression equations estimated by Landau and (8), the estimated coefficients of G/Y in Table 1 show a pattern that seems remarkably similar to the one reported by him. It is difficult, however, to share his conclusion that "These results are consistent with a pro-free market view that—within the market economies—a growth of government hurts economic growth" (p. 790). The results he obtained apparently rest on a misspecification of the growth equation. As the discussion in Section II indicates, the appropriate variables to investigate whether "growth of government hurts economic growth" are $\dot{G}$ and/or $\dot{G}(G/Y)$, and not G/Y, whether one uses conventional reasoning of neoclassical growth models, or adopts the augmented models proposed in Section II. Estimates obtained by using G/Y seem to throw little light on the issue, and the more defensible specifications in (5), (6), and (7) yield results that are exactly the opposite of those obtained by using G/Y as a regressor in the growth equation.[13]

Besides the use of G/Y as a regressor, Landau's specification has some other features that differ from (8), and one could ask how important those might be. For example, his equation does not have an I/Y term, but has instead a human capital investment term based on school enrollments. It seems difficult to justify exclusion of I/Y or inclusion of human capital measures premised on *current* school enrollment rates of children as determinants of economic growth. Also, more inclusive specifications with regional "dummies" of the kind employed by Landau yield the same pattern of estimates as reported in Table 1. His per capita output growth equation can be regarded as a special case of the aggregate growth equation in which the coefficient of $\dot{P}$ is constrained to unity. In short, therefore, it appears likely that the negative parameter estimates for the

[12] The discussion in this section is centered on government variables that are of direct relevance to the study. Several other estimates also indicate some interesting patterns. For instance, the coefficients of the population growth variable for the *LDC* group in 1970–80 are lower than for the full sample in that period and are also lower than the corresponding estimates for 1960–70.

[13] A study of the connection between G/Y and *GNP* per capita seems close to an assessment of the hypothesis advanced by Adolph Wagner (1890) regarding the scale of state activity. However, as might be expected, much of the literature on Wagner's hypothesis suggests that the flow of causality is *from* economic development *to* G/Y, and not from G/Y to an increase of *GNP* as implied in specifications like equation (8).

TABLE 2–ILLUSTRATIVE RESULTS (t-STATISTICS) FOR DAVIDSON AND MACKINNON'S J-TEST[a]

	1960–70		1970–80	
	LDCs and *DCs*	*LDCs*	*LDCs* and *DCs*	*LDCs*
H_0: (8) is true H_1: (6) is true	4.82	4.20	6.90	5.76
H_0: (6) is true H_1: (8) is true	−0.96	−0.60	0.96	2.00

[a]The number of degrees of freedom in the full sample (*LDCs* and *DCs*) is 110, and that in the *LDC* subsample 89. The critical 1 percent (two-tailed) t-statistic in both cases is around 2.65.

government variable in Landau's work arise mainly due to the inappropriate use of G/Y as a regressor.

The main conclusion reached by Rubinson is similar to that suggested by Table 1. However, he also used a variable similar to G/Y as one regressor, and his methodology is different in several other respects too. While his conclusion is supported by this study, a direct or useful comparison with his work seems difficult.

IV. Some Model Specification Tests

It has been claimed in the foregoing discussion that equations (5), (6), and (7) are superior to specifications that resemble (8). Some theoretical reasons for that claim have already been given; equation (8) is hard to derive from production models of the kind usually considered plausible in such cases, and it can yield no information about the externality effect of government size or about the intersectoral productivity differential, two major mechanisms through which government size may affect economic performance.

Apart from considerations of the underlying economic theory or the informational content of the parameters, it is also possible to compare the models in a statistical sense. The simplest criterion for such a comparison would be the "goodness of fit" of the models to the data. It is easy to see from Table 1 that criterion clearly favors (6) and (7), and probably (5), over (8); in every single case R^2 for (8) is lower, often by a large margin, than for the other equations even though the number of regressors in (6), (7), and (8) is the same.

Several direct and formal tests of model specification in such nonnested cases have been proposed by Davidson and MacKinnon. Their "J-test" can be easily used since the models are linear. The test procedure is detailed in the cited work and is mentioned here in the barest outline. Taking the case of two linear models, let the competing specifications be

$$y = X_1\beta_1, \tag{9}$$

$$y = X_2\beta_2, \tag{10}$$

where y is a vector of observations on the dependent variable, X_1 and X_2 are the matrices of observations on x's in the two cases, and β_1 and β_2 are the respective parameter vectors. The models are nonnested in the sense that neither X matrix is a subset of the other. The hypothesis H_0: (9) is "true" (and 10) is false); and H_1: (10) is "true" (and (9) is false) can be directly investigated by obtaining predicted values of y from (10) and, adding them as observations on an additional regressor in (9), testing the significance of its coefficient through a straightforward t-test. As Davidson and MacKinnon point out, one should also test the reverse hypothesis for a more secure conclusion.

Applying the J-test to (6) and (8), which differ only in terms of one variable, it is seen that the predominance of the test statistics clearly favors (6) over (8); the hypothesis that (8) is true is rejected in every case at the

1 percent level, while the hypothesis that (6) is the true model is accepted in most cases. Table 2 illustrates the pattern of the t-statistics obtained. Similar tests were also conducted to compare (7) and (8), and the predominance of evidence favors (7) over (8).[14]

It seems fair to conclude that whether one considers the underlying economic reasoning, the information content of the parameters, fit with the data, or more formal tests, equations (6) and (7) seem preferable to those like (8).[15]

V. Evidence from Time-Series Data

With some rare exceptions, almost every study of the impact of government size on growth is based on cross-section data averaged over various periods. An obvious reason is the paucity of usable time-series data except possibly for a few industrialized countries. However, although useful for obtaining broad indications, such cross-section models imply strong parametric restrictions across countries that often differ greatly in terms of their economic structures. It seems useful to make a beginning toward considering evidence based on time-series data for as many countries as feasible.

Advantage is taken here of the opportunity provided by the publication of the Summers-Heston data on a time-series basis from 1960 through 1980 for 115 countries.[16] Since the data set yields only 20 observations for each country, one cannot expect very sharp estimates, and considerable caution is needed in interpreting the estimated parameters. Nevertheless, it should be useful to make a broad judgment on the basis of the overall pattern revealed by the estimates.

The preferred equations (6) and (7) are estimated for each of the 115 countries. Estimation was done with ordinary least squares (OLS) and also on the premise of a first-order autoregressive disturbance term ($AR1$). Based on OLS estimates, regression F-statistics are found to be statistically significant at least at the 10 percent level in 70 cases. Table 3 contains detailed estimates for both (6) and (7) for these 70 cases. The $AR1$ estimates are reported wherever the autoregressive parameter is statistically significant at the 10 percent level; OLS estimates are reported in other cases.[17]

Panel A in Table 4 summarizes the position revealed by the estimates. Considering the entire sample, the estimated coefficient of $\dot{G}$ in (6) is positive in 100 of the 115 cases, and the coefficient of $\dot{G}(G/Y)$ in (7) is positive in 98 cases; for both, 55 percent to 60 percent of the positive coefficients are statistically significant; only in one of the 115 cases is either estimate significantly negative. Considering the more interesting group of 70 significant regressions, positive coefficients for $\dot{G}$ in equation (6) and for $G(G/Y)$ in (7) are observed in 66 and 65 cases, respectively; nearly three-fourths of these are statistically significant; and, again, there is only one case of a significant negative coefficient.

Therefore, despite the obvious limitations of the data, especially the small number of observations for each country, estimates from time-series data for this large group of countries convey the same message as that based on cross-section models, namely, government size exercises a statistically significant posi-

[14] Note that the Davidson-MacKinnon tests are designed for nonnested models; nested models can be compared directly. Details of the test results, in addition to those given in the text, are available from the author. See also David Spencer and Hamzaid Yahya (1984) for an interesting application of the Davidson-MacKinnon tests.

[15] While this section focuses on the Davidson-MacKinnon tests, some other specificational aspects, which are common to (5), (6), (7), and (8), were also looked at. The possibility of heteroscedasticity of the error term and/or existence of a "simultaneous equations bias" was investigated by using the test proposed by Halbert White (1980, p. 824). The hypothesis of homoscedasticity and no specification error is easily maintained for 1970–80, but not for 1960–70. However, since the focus is on parameter signs, it seems unlikely that the main conclusions are affected by this aspect.

[16] There are 63 countries for which data for 1950–59 are also available. Their results for 1950–80 are similar to those in Table 3. Because of the lower reliability and comparability of data for years prior to 1960, it seems better to focus on 1960–80 despite the limited number of observations.

[17] The $AR1$ and OLS estimates are very similar in most cases. Detailed estimates, in addition to those given in Table 3 are available on request.

TABLE 3—ESTIMATED COEFFICIENTS OF GOVERNMENT VARIABLES IN EQUATIONS (6) AND (7) DERIVED FROM TIME-SERIES DATA FOR INDIVIDUAL COUNTRIES: 1960–80[a]

Country	Method	Equation (6) ($\hat{\theta}$)	Equation (7) ($\delta' + C_G$)	Country	Method	Equation (6) ($\hat{\theta}$)	Equation (7) ($\delta' + C_G$)
Afghanistan	*OLS*	0.337	2.484	Ivory Coast	*AR*1	0.466	2.313
		(3.59)	(3.63)			(4.85)	(4.00)
Algeria	*OLS*	0.471	2.725	Jamaica	*OLS*	0.044	0.048
		(3.44)	(3.87)			(0.47)	(0.10)
Australia	*OLS*	−0.166	−1.499	Jordan	*OLS*	0.500	1.376
		(−2.44)	(−2.42)			(3.03)	(2.72)
Austria	*OLS*	−0.130	−0.818	Korea (South)	*OLS*	0.599	4.363
		(−1.33)	(−1.26)			(4.64)	(4.47)
Barbados	*OLS*	0.213	1.199	Madagascar	*OLS*	0.378	1.522
		(2.79)	(2.63)			(7.17)	(6.68)
Belgium	*OLS*	0.109	1.043	Mali	*AR*1	0.434	1.880
		(1.08)	(1.06)			(4.34)	(4.14)
Bolivia	*OLS*	0.177	1.187	Malta	*AR*1	0.233	0.977
		(2.88)	(2.94)			(3.29)	(3.22)
Brazil	*AR*1	0.304	1.644	Mauritius	*OLS*	0.281	1.450
		(2.49)	(2.18)			(2.57)	(2.66)
Burma	*OLS*	0.745	3.559	Morocco	*OLS*	0.172	0.714
		(15.66)	(15.74)			(1.78)	(1.70)
Burundi	*OLS*	0.162	1.404	Mozambique	*OLS*	0.598	3.028
		(1.73)	(2.32)			(7.75)	(7.76)
Cameroon	*OLS*	0.221	0.853	Nepal	*OLS*	0.171	0.683
		(2.79)	(2.52)			(3.10)	(2.89)
Chile	*OLS*	0.219	1.769	Netherlands	*OLS*	0.021	0.284
		(1.08)	(1.26)			(0.17)	(0.26)
Columbia	*AR*1	0.146	1.366	New Zealand	*OLS*	0.122	0.904
		(1.98)	(1.94)			(1.42)	(1.38)
Costa Rica	*OLS*	0.191	1.058	Niger	*AR*1	0.592	2.598
		(2.00)	(1.99)			(3.91)	(3.36)
Denmark	*OLS*	0.286	1.618	Pakistan	*OLS*	0.243	1.076
		(1.32)	(1.30)			(4.83)	(5.43)
Dom. Republic	*OLS*	0.168	1.074	Panama	*AR*1	−0.035	−0.229
		(2.35)	(2.87)			(−0.39)	(−0.45)
Ecuador	*OLS*	0.307	1.207	Papua New Guinea	*OLS*	0.402	1.192
		(3.32)	(2.97)			(2.33)	(2.37)
Egypt	*OLS*	0.155	0.791	Paraguay	*OLS*	0.032	0.150
		(1.45)	(1.70)			(0.85)	(0.61)
El Salvador	*OLS*	0.039	0.057	Peru	*AR*1	0.056	0.296
		(0.18)	(0.05)			(0.85)	(0.79)
France	*OLS*	0.193	1.586	Philippines	*OLS*	0.077	0.478
		(1.91)	(1.94)			(0.92)	(1.06)
Gambia	*AR*1	0.382	1.622	Portugal	*OLS*	0.252	1.829
		(2.43)	(2.43)			(1.32)	(1.34)
Germany (FRG)	*OLS*	−0.093	−0.764	Rwanda	*OLS*	0.577	2.867
		(−0.63)	(−0.71)			(4.15)	(3.18)
Guatemala	*OLS*	0.025	0.183	Senegal	*AR*1	0.743	3.243
		(0.42)	(0.43)			(5.11)	(4.96)
Guinea	*OLS*	0.234	1.157	Sierra Leone	*OLS*	0.158	0.897
		(2.82)	(2.40)			(1.41)	(1.33)
Guyana	*OLS*	0.330	0.920	Singapore	*OLS*	0.372	2.396
		(4.66)	(2.83)			(3.76)	(3.58)
Haiti	*OLS*	0.340	1.637	South Africa	*OLS*	0.358	2.182
		(4.54)	(4.04)			(3.02)	(2.89)
Hong Kong	*OLS*	0.076	0.643	Sri Lanka	*OLS*	0.510	1.780
		(2.51)	(2.40)			(6.30)	(7.96)
Iceland	*OLS*	0.462	5.157	Suriname	*AR*1	0.484	2.244
		(2.64)	(2.31)			(3.57)	(3.62)
India	*AR*1	0.248	1.451	Swaziland	*OLS*	0.339	1.600
		(6.03)	(6.15)			(2.94)	(3.49)
Indonesia	*OLS*	0.107	0.616	Switzerland	*OLS*	0.092	1.023
		(2.71)	(2.62)			(0.94)	(0.88)
Iran	*OLS*	0.479	2.507	Turkey	*OLS*	0.138	0.460
		(3.97)	(3.60)			(2.66)	(2.38)
Italy	*OLS*	−0.195	−1.395	Uganda	*AR*1	0.175	1.094
		(−1.41)	(−1.32)			(2.77)	(2.88)

(continued)

TABLE 3—Continued

Country	Method	Equation (6) ($\hat{\theta}$)	Equation (7) ($\delta' + C_G$)	Country	Method	Equation (6) ($\hat{\theta}$)	Equation (7) ($\delta' + C_G$)
U. K.	*OLS*	0.146	0.607	Uruguay	*AR*1	0.335	1.706
		(1.00)	(0.91)			(4.11)	(4.25)
U.S.A.	*OLS*	0.009	−0.053	Zaire	*OLS*	0.422	1.421
		(0.07)	(−0.06)			(5.23)	(5.11)
Upper Volta	*OLS*	0.337	1.575	Zambia	*AR*1	0.415	1.361
		(6.69)	(7.90)			(2.95)	(2.73)

[a]The t-statistics are shown in parentheses. Included countries are those for which the regression F-statistic is significant at least at the 10 percent level. *AR*1 estimates are based on *AUTOREG* procedure of Statistical Analysis System. The *AR*1 estimates are reported for all cases in which the autoregressive parameter is statistically significant at least at the 10 percent level.

TABLE 4—SUMMARY OF TIME-SERIES RESULTS

	Entire Sample		Table 3 Sample	
	Equation (6)	Equation (7)	Equation (6)	Equation (7)
A. Signs of the Coefficients of Government Variables in Equations (6) and (7)				
Positive coefficient, significant at least at the 5 percent level	48	47	43	44
Positive coefficient, significant at 10 percent, but not at 5 percent or better	8	7	5	4
Positive coefficient, not significant even at 10 percent	44	44	18	17
Total	100	98	66	65
Negative coefficient, significant at least at the 5 percent level	1	1	1	1
Negative coefficient, significant at 10 percent, but not at 5 percent or better	0	0	0	0
Negative coefficient, not significant even at 10 percent	14	16	3	4
Total	15	17	4	5
Total Sample Size	115	115	70	70

B. Some Simple Correlations between Coefficients of Government Variables in Equations (6) and (7) and *GDP* per capita in 1970 (*RY*70) and the ratio G/Y: Table 3; Sample ($N = 70$)

	Coefficient in Equation (6)	Coefficient in Equation (7)	*RY*70
Coefficient in Equation (7)	0.88[a]		
*RY*70	−0.48[a]	−0.32[a]	
G/Y	0.48[a]	0.10	−0.48[a]

[a]Statistically significant at least at the 5 percent level.

tive effect on economic performance in most cases.[18]

Rubinson reported that the positive effect of government on growth is stronger in poorer *LDC*s. It is of some interest to make a judgment on that aspect from the results obtained in this study. The cross-section estimates in Table 1 appear somewhat uninformative on that issue. Since the time-series data yield the overall effect of government size on growth for each country, it is easy to see if the estimated effect correlates with the country's income. Panel B in Table 4 provides some simple correlations of the coefficients of both $\dot{G}$ and $\dot{G}(G/Y)$ with real *GDP* per capita for 1970. The correlations are negative, have a modest magnitude, and are statistically significant. Therefore, although one need not rely on the "dependency" argument to explain the results, there is some evidence to suggest that the positive effect of government on growth is typically stronger at lower income levels. It is difficult to say whether this is simply some kind of a scale effect. The simple correlations between G/Y and the two coefficients, also reported in Panel B of Table 4, are positive and do not quite support the "scale-effect" hypothesis.

VI. Concluding Remarks

Several well-known caveats are needed in interpreting the statistical estimates obtained in such cross-country studies. Nevertheless, the issue is important, and it appears useful to investigate at least the direction of the effect of government size on economic performance, using a reasonably defensible theoretical framework and good, internationally comparable data for as many countries as one can.

This work employs a framework that, although simple, has considerable appeal in terms of modeling the externality effect of government size, relative factor productivity in the government and nongovernment sectors, and the overall impact of government size. Besides being more informative than some similar conventional specifications, the models used are derived from plausible production relations, fit the data better, and are favored by formal statistical tests. The number of countries included is 115, which is perhaps the largest cross-country set ever used to study the issue. The data are internationally comparable and probably the best one can get for such a large group of countries. The parametric rigors of cross-section models are sought to be softened somewhat by obtaining results from time-series data also, even though the number of observations in each case is limited.

The main result is that it is difficult not to conclude that government size has a positive effect on economic performance and growth, and the conclusion appears to apply in a vast majority of the settings considered. Even more interesting seems to be nearly equally pervasive indication of a positive externality effect of government size on the rest of the economy. It is also possible to infer from the cross-section evidence that relative factor productivity was higher in government sector than in the rest of the economy at least during the 1960's. At a more detailed level, three tentative results are discernible; the positive externality effect of government may have increased over the 1970's; relative factor productivity in the government sector could have declined during that period; and the positive effect of government size on growth could well be stronger in lower-income contexts.

[18] Once again, it is worth noting that results from time-series data are best used to get a feel for the overall pattern. Although the parameter estimates are good and plausible for a very large number of countries, there are obvious weaknesses in several cases.

REFERENCES

Afxentiou, P. C., "Economic Development and the Public Sector: An Evaluation," *Atlantic Economic Journal*, December 1982, *10*, 32–38.

Davidson, Russell and MacKinnon, James G., "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, May 1981, *49*, 781–93.

Feder, Gershon, "On Exports and Economic

Growth," *Journal of Development Economics*, February/April 1983, *12*, 59–73.

Kravis, Irving, Heston, Alan and Summers, Robert, *World Product and Income International Comparisons of Real Gross Product*, Baltimore: Johns Hopkins University Press, 1982.

Landau, Daniel, "Government Expenditure and Economic Growth: A Cross-Country Study," *Southern Economic Journal*, January 1983, *49*, 783–92.

Rubinson, Richard, "Dependency, Government Revenue, and Economic Growth, 1955–70," *Studies in Comparative International Development*, Summer 1977, *12*, 3–28.

Spencer, David E. and Yahya, Hamzaid B., "Financial Development and the Demand for Money," Department of Economics Working Paper, Washington State University, 1984.

Summers, Robert and Heston, Alan, "Improved International Comparisons of Real Product and its Composition: 1950-80," *Review of Income and Wealth*, June 1984, *30*, 207–262.

______, **Kravis, Irving B. and Heston, Alan,** "International Comparisons of Real Product and its Composition: 1950-77," *Review of Income and Wealth*, March 1980, *26*, 19–66.

Wagner, Adolph, *Finanzwissenchaft*, Leipzig, 1890.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

World Bank, *World Tables*, 3rd ed., Baltimore: Johns Hopkins University Press, 1984.

A Macroeconomic Model with Auction Markets and Nominal Contracts

By RICHARD CANTOR*

This paper analyzes a macroeconomic model in which labor market participants are distributed endogenously between an auction market and nominal contracting sector. The microeconomic choice between arrangements embodies a tradeoff of production efficiency and efficiency in the allocation of risk. Conditions for the coexistence of the two sectors are derived. The contracting sector provides a channel for activist monetary policy familiar to the nominal contracting literature. The major policy implications of this literature are shown to be robust to the case in which the distribution of agents across sectors responds to alternative monetary rules.

Stanley Fischer (1977) and Edmund Phelps and John Taylor (1977) have argued that economic agents bind themselves to nominal contracts which make activist monetary policy possible and perhaps desirable. This literature has been criticized (see, for example, Robert Barro, 1977) for its failure to specify the implied optimizing behavior which gives rise to these contracts. In fact, according to the *ad hoc* welfare criterion adopted by Fischer, auction markets for labor services clearly dominate nominal contracts. In the optimizing model which follows, it is shown that the use of nominal contracts increases with policy activism because such policies increase the *ex post* efficiency of contracts and make them more attractive *ex ante*.[1]

In developing some micro foundations for macroeconomic models with nominal contracts, the first step is to describe an environment in which agents rationally demand nominal contracts. The literature has long recognized that fully contingent contracts dominate those with restricted contingency sets; however, these first-best arrangements imply prohibitively high writing, monitoring, and enforcement costs and can be ruled out a priori. Therefore, contracts with restricted contingency sets may be adopted, and the aggregate price level is a likely element of the included set.

In a simple model economy with real and monetary shocks, the optimal price contingent (nominal) contract structure is derived in an implicit contracts framework. The benefits of contracting, in this model, are derived from efficient risk-sharing arrangements between firms and unions (or workers). Different unions place different valuations on these contracts because their attitudes toward risk vary.

Firms and unions are not limited to this type of labor market arrangement. They may foresake price contingent contracts (*PCCs*) in order to participate in a full-information auction market for labor services (*AM*). Their choice reflects a tradeoff between production efficiency (information and flexibility *ex post*) vs. the value of precommitment (insurance *ex ante*).

The model economy is described in Section I, and the *PCCs* and *AM* are analyzed in Section II. The supply, demand, and equilibrium price of contracts are analyzed in Section III. Monetary policy is discussed briefly in Section IV and is followed by some concluding remarks.

*Department of Economics, Ohio State University, Columbus, OH, 43210. I am grateful for comments received on a earlier draft from Stephen McCafferty, Robert Driskill, three anonymous referees, and my thesis advisors, Edi Karni and Louis Maccini.

[1] Using JoAnna Gray's (1978) model of contract duration, Matthew Canzoneri (1980), Gary Fethke and Andrew Policano (1981), and John McCallum (1983) have shown that activist policy rules increase contract length and retain their effectiveness. However, since Gray's model lacks a natural welfare measure and does not offer agents an auction market alternative, the robustness of this "effectiveness" result needs to be demonstrated.

I. The Model Economy

Assume there exists an infinity of identical firms and, for convenience, let each firm be associated with a point along the unit interval. Further, assume there exists a continuum of unions, one associated with each point along the unit interval. There is exactly one firm for each union; however, an alternative interpretation is that each firm employs the same fraction of the labor force. The bargaining environment is fully competitive in the sense that any union is free to supply labor services to any one firm.

Each firm produces a homogeneous commodity by employing the labor services n (hours) of one union. The production process is governed by the following function,

$$(1) \qquad y = kf(n); \qquad f' > 0,\ f'' < 0,$$
$$\lim_{n\to\infty} f'(n) = 0 \qquad \lim_{n\to 0} f'(n) = \infty,$$

where k is a random aggregate productivity shock defined over the interval $(0,\infty)$. Output is sold at a competitive price p which is the same as the aggregate price level.

Firms are assumed to be risk neutral and maximize expected real profits. Real profits π are defined by

$$(2) \qquad \pi = y - wn/p,$$

where w is the nominal wage. Unions' preferences over income and leisure are assumed to be represented by the following utility function,

$$(3) \quad u = u(wn/p - g(n))$$
$$g' > 0, g'' > 0, u' > 0, u'' \leq 0.$$

The concavity of $u(\cdot)$ embodies risk aversion and convexity of $g(\cdot)$ implies that the marginal disutility of labor increases at the margin. The class of utility functions (3) includes all those for which the income elasticity of labor supplied is zero. Restricting attention to this class of utility functions greatly simplifies the analysis, but the main results should be robust to generalizations along these lines.

Unions differ only by degree of absolute risk aversion and are ordered along the unit interval such that the first union is infinitely risk averse, the last is risk neutral, and the degree of risk aversion varies continuously along the interval. This restriction can be captured by a parameter q characterizing a Diamond-Stiglitz-type (1974) utility function such that q indexes the unions' degrees of risk aversion. That is,

$$(4) \qquad u = u(wn/p - g(n); q), \quad \forall q \epsilon [0,1];$$

$$(5) \qquad \frac{d}{dq}\left[\frac{-u''(Ez;q)}{u'(Ez;q)}\right] < 0, \qquad \forall Ez;$$

$$(6) \quad \frac{-u''(Ez;0)}{u'(Ez;0)} = \infty; \frac{-u''(Ez;1)}{u'(Ez;1)} = 0; \qquad \forall Ez,$$

where z is a random variable defined below.

In general, the price level is functionally related to all the real and nominal disturbances in the macroeconomic system; therefore, p and k are jointly distributed. In particular, p and k are generally negatively related since a positive productivity shock increases aggregate supply which reduces the price level if the money stock is held constant.

There is a nonnegative monetary disturbance e that consists of both money supply and demand (velocity) shocks and is independent of k. Monetary equilibrium is described by

$$(7) \qquad py = me,$$

where y is aggregate output and m is the money supply controlled by the authorities and both are expressed in per union terms.

It is useful at this point to define $E(k|p)$ and $V(k|p)$ as the conditional expectation and variance of k given p, respectively. I assume that these functions exist and are known by the firms and unions. The joint distribution of k and p depends upon the underlying distributions of k and e and the labor market arrangements chosen by firms and unions. In an equilibrium with two

labor markets, average per union output is given by

$$(8) \qquad y = N_P y_P + (1 - N_P) y_A,$$

where y_P and y_A are outputs per *PCC* and *AM* firm, respectively, and N_P is the fraction of firms that contract for labor.

II. The Expected Benefits of Labor Market Participation

Ex ante, agents must choose whether to contract or to wait until k and p are realized and the *AM* opens. *Ex post* contracting is meaningless because the opportunity for risk sharing is past once the state of the world is realized. Since each firm employs only one union and there are an identical number of firms and unions, then, by construction, *AM* wages and employment per firm are independent of the size of the *AM* sector. Therefore, agents can calculate *ex ante* their expected rewards from *AM* participation. Agents competitively demand and supply contracts subject to the constraint that the expected contractual rewards are at least as great as the rewards available in the *AM*.

A. *The Expected AM Rewards*

In the *AM*, agents competitively demand and supply labor with complete knowledge of the realized values of p and k. The equilibrium wage serves to clear the *AM* for labor services and implies that *AM* unions receive their actual marginal products, and that equilibrium employment equates the marginal product $kf'(n)$ to the union's marginal disutility of labor $g'(n)$. This implies that *AM* employment can be expressed as a strictly increasing function of the productivity shock:[2]

$$(9) \qquad n_A = \phi(k).$$

Since *AM* unions receive their marginal products, their incomes are

$$(10) \qquad (wn/p)_A = kf'(\phi(k))\phi(k).$$

If the stochastic distribution of k is known, the *ex ante* expectations of utility and profits can be calculated directly. Substituting from (9) and (10) into (4) and taking expectations yields expected *AM*, union utility,

$$(11) \qquad Eu_A = E[u(z;q)],$$

where "net income" z is given by

$$z = kf'(\phi(k))\phi(k) - g(\phi(k)).$$

Substituting from (9) and (10) into the definition of profits and taking expectations yields expected *AM*, firm profits,

$$(12) \quad E\pi_A = E[kf(\phi(k)) - kf'(\phi(k))\phi(k)].$$

Agents choose their preferred labor market arrangement based upon a comparison of the benefits from *AM* participation with those from contracting. Equation (11) indicates that unions value *AM* income differently depending upon their differences in risk aversion as indexed by q; however, the model presented here in no way requires interpersonal comparisons of utility. Equation (12) implies that expected profits are the same for all *AM* firms.

B. *Expected PCC Rewards*

Prior to the realization of k and p and the opening of the *AM*, a firm-union pair has the option of entering into a binding contract which specifies wages and employment levels for each possible realization of the aggregate price level. Certainly a larger contingency set including realizations of k would be desirable; however, such contracts are ruled out by assumption.[3] Therefore, firms are limited

[2] The unions degree of risk aversion does not affect its *AM* labor supply because that decision is made *ex post*. The subscripts A and P refer to the *AM* and *PCC* values, respectively.

[3] A similar contract has been analyzed by Costas Azariadis (1978), but his real shock did not effect workers' marginal productivity. The observed absence of linkages to real aggregate variables has not been ex-

to PCCs of the following structure,

$$\delta = \{w(p), n(p)\}.$$

The linkage of employment to price may seem odd; however, since w is also a function of p, one may interpret δ as linking n to w.[4]

Optimal implicit contracts are usually derived in models with identical unions and under the restriction that unions are guaranteed a market-determined expected utility level $\overline{U}$. In this section, the optimal PCC is derived under these assumptions. Then it is shown that this contract is also optimal in an economy with heterogeneous unions and an AM alternative to contracts.

The optimal contract solves

$$\underset{\{\delta\}}{\text{Max}}\, E[kf(n) - wn/p]$$

subject to $E[u(wn/p - g(n))] = \overline{U}$.

Contingencies cannot be made directly to k, but realizations of p may reveal information about probable realizations of k.

The first-order conditions characterizing the optimal PCC are

$$(13)\quad 1/\lambda = u'(wn/p - g(n)), \quad \forall p,$$

$$(14)\quad E(k|p)f'(n) = g'(n), \quad \forall p,$$

$$(15)\quad E[u(wn/p - g(n))] = \overline{U},$$

where λ is the state-independent Lagrangian multiplier.

Equation (13) is the familiar Borch-Arrow optimal insurance condition which implies that the risk-averse union's utility is kept constant across all states. The assumed specification of the utility function therefore implies that net income is constant,

$$(16)\quad T = wn/p - g(n), \quad \forall p,$$

where T is a constant determined by equation (15).

Equation (14) determines PCC employment such that the conditional expectation of the marginal product is equated to the marginal disutility of labor. This implies the PCC employment function is

$$(17)\quad n_P = \phi(E(k|p)) \equiv \hat{\phi}.$$

That is, the function has the same curvature as the AM employment function but its argument is $E(k|p)$.

So far, only the case in which unions are identical has been considered. In fact, differences among unions with respect to risk aversion do not affect the structure of the contracts offered. Since almost all unions are risk averse, firms are led to offer fixed income contracts only. The optimal employment and wage functions are independent of the union's degree of risk aversion. All PCC unions receive this income T and will enjoy the following utility

$$(18)\quad EU_P = u(T; q).$$

By substitution from (16) and (17) into the definition of expected profits, one obtains,

$$(19)\quad E\pi_P = E[kf(\phi(E[k|p])) - g(\phi(E[k|p])) - T].$$

All PCC firms expect to earn the same profits.

If firms are competitive in the sense that they offer contracts which provide unions with a market-determined income T, then firms derive no additional benefit from contracting with particularly risk-averse workers.

---

plained by the contract literature. I rely on the *ad hoc* "prohibitive cost" argument in order to derive from an optimizing model a nominal contract comparable to Fischer's. In a personal communication, Stephen McCafferty has pointed out that, in the current model, observation of a single AM worker's wage reveals k. However, a simple model in which workers vary according to skill would alter none of the results, but would require knowledge of the AM sector's skill composition and every worker's wage in order to deduce k.

[4] If firms have *ex post* knowledge of k but workers do not, then δ may be dominated by a contract structure in which firms can unilaterally control employment *ex post*. This possibility has been excluded for analytical simplicity but would seem not to affect the main results. Such a contract structure has been analyzed by Canzoneri and Anne Sibert (1984) and was shown to condition employment levels on p in a similar manner as the optimal contract δ (see equation (17), below).

This stipulation is equivalent to the assumption that competitive suppliers of any product cannot appropriate the consumers' surplus of inframarginal demanders; that is, competitive firms do not practice price discrimination.

In general, unions prefer contracts which offer large values of T and firms prefer small T's. Unions that are more risk averse (represented by small values of q) are more likely to prefer a contractually fixed T relative to the alternative AM income than are less risk-averse unions.

III. Demand, Supply, and the Equilibrium Contract Value

Ex ante, firms competitively supply contracts which accord unions an equilibrium level of nonstochastic income T^*. Those unions and firms which do not contract for labor, supply, and demand labor services at the competitive, full-information (*ex post*) AM wage. Since T^* equates supply and demand for contracts, it implies a particular equilibrium fraction of firms contracting for labor N_P^*. This section discusses the determinants of the demand and supply for contracts and the conditions that imply the coexistence of both arrangements in equilibrium.

The variable T may be thought of as the price of a contract; however, T is paid by the suppliers (firms) and received by the demanders (unions). I characterize the demand for contracts as follows: define $D(\tilde{q})$ as the reservation T (demand price) for a union with risk aversion $\tilde{q}$; that is, $D(\tilde{q})$ is the minimum contractual net income sufficient to induce a union $\tilde{q}$ to contract.

PROPOSITION 1: *The demand price for contracts as measured by $D(q)$ has the following properties*:

(*i*) $D'(q) > 0$,

(*ii*) $D(1) = Ez$, *and*

(*iii*) $D(0) \geq \mathbf{z}$, *where* $\mathbf{z}$ *is the lower bound of the distribution of* z (*an AM union's stochastic net income*).

PROOF:

The definition of $D(q)$ implies that it is an implicit function which equates contractual utility to expected AM utility for every q,

$$(20) \qquad u(D(q);q) = E[u(z;q)], \qquad \forall q.$$

A union's risk premium, $Ez - D(q)$ is the amount of expected income it would forego in order to receive a nonstochastic income. According to Peter Diamond's and Joseph Stiglitz's Theorem 3, the risk premium is strictly decreasing with q and, therefore, $D'(q) > 0$. If $q = 1$, then the union is by assumption risk neutral and the risk premium is zero, implying $D(1) = Ez$. The function $D(q)$ must be greater then or equal to $\mathbf{z}$ for any q since, were it not true, a union would choose a contractual income that is less than AM income in every possible state of the world.

A risk-neutral union would choose to contract only if T were at least as great as the expected AM income Ez. The reservation T declines with the union's degree of risk aversion. Intuition suggests that an infinitely risk-averse union ($q = 0$) would be willing to accept a contract offering "*epsilon*-more" income than $\mathbf{z}$. This assertion can be easily demonstrated at least for the special case in which z is uniformly distributed and the $u(\cdot;0)$ exhibits constant absolute risk aversion.

The function $D(q)$ defines a demand curve for contracts. The supply curve is derived by comparing $E\pi_A$ with $E\pi_P$. Since all firms are identical, they share the same reservation T at which they are indifferent between AM participation and PCCs. Define this reservation contract price as S which implicitly solves the equation $E\pi_A - E\pi_P = 0$. Substituting from (12) and (19) into this expression yields,

$$(21) \qquad Ez - S = E\big[kf(\phi(k)) - g(\phi(k)) - kf(\phi(E(k|p))) + g(\phi(E(k|p)))\big].$$

The rather foreboding expression on the right-hand side of (20) can be interpreted as the value of the omitted contingency and the greater production efficiency of the AM. The terms $kf(\cdot) - g(\cdot)$ represent the social gain from production (the real value of output minus the social labor cost). The right-hand

side is greater than zero because *PCC*s cannot link employment directly to k and therefore imply states of inefficient employment. Firms would not willingly bear this cost of *PCC*s unless they received lower labor costs ($S < Ez$). The supply price S is characterized more fully in the following proposition.

PROPOSITION 2: *For small risks, the supply price S of a contract is given by*

$$S = Ez - \tfrac{1}{2}E[V(k|p)f'(\hat{\phi})\hat{\phi}], \tag{22}$$

where $V(k|p)$ is the variance of k conditional on p and $\hat{\phi} = \phi(E(k|p))$.

PROOF:

For any realization of $p = \bar{p}$, define $\bar{k} = E[k|p = \bar{p}]$ and $\Psi(k, \bar{p})$ as

$$\Psi(k, \bar{p}) = kf(\phi(k)) - g(\phi(k)) - kf(\phi(\bar{k})) + g(\phi(\bar{k})).$$

Notice that the unconditional expectation of $\Psi(k, p)$ is $Ez - S$. Recalling that $kf'(\phi(k)) \equiv g'(\phi(k))$ holds as an identity, the second-order approximation of $\Psi(k, \bar{p})$ around $k = \bar{k}$ is given by

$$\Psi(k, \bar{p}) = (k - \bar{k})^2 \tfrac{1}{2} f'(\phi(\bar{k}))\phi'(\bar{k}).$$

The conditional expectation of $\Psi(k, \bar{p})$ taken over all values of k yields

$$E[\Psi(k, \bar{p})|p = \bar{p}] = V(k|p = \bar{p})f'(\phi(\bar{k}))\phi'(\bar{k}).$$

The unconditional expectation of $\Psi(k, p)$ yields the expression for $Ez - S$ given in equation (22).

The last term on the right-hand side of (22) represents the implicit cost of contracting for firms. Since *PCC*s give rise to inefficient *ex post* employment allocations, expected profits would be greater in the *AM* if labor costs were the same. The size of this implicit cost depends on two factors, $E[V(k|p)]$ and $f'\phi'$. The first term represents the quality of p as a signal for k because *PCC*s reply on price level fluctuations to reveal movements of k. The second term represents the overall importance of information about k. The variable ϕ' measures the responsiveness with which employment should optimally vary with k, and f' translates employment movements to output movements.

Although the supply price of an individual firm can be read from equation (22), the aggregate supply curve is more complicated because the $E[V(k|p)]$ is a function of N_P, the fraction of the economy engaged in contracting. Under the assumptions listed in the Lemma below, it is clear that the aggregate supply price of contracts increases with N_P. The intuition behind this result suggests that it may be robust to relaxation of these assumptions.

LEMMA: *Given the following assumptions:*

(a) *$f(n) = n^a$ and $g(n) = n^b$, where $0 < a < 1$ and $b > 1$;*

(b) *$\log\ k \sim N(0, \sigma_k^2)$ and $\log\ e \sim N(0, \sigma_e^2)$; and*

(c) *a fixed money stock with monetary equilibrium given by (7). Then, for small risks, if the ratio of monetary to real "noise," σ_e^2/σ_k^2, is positive, the $V(k|p)$ is strictly increasing with N_P and σ_e^2/σ_k^2. If $\sigma_e^2/\sigma_k = 0$, then $V(k|p) = 0$.*

This Lemma is proven by directly calculating the joint distribution of p and k as shown in my 1983 working paper.

This result can be explained as follows. When there is a productivity increase (decrease), output in both labor market sectors increases (falls). Since *AM* agents can use their knowledge of k, output in the *AM* will increase (decrease) more than in the contractual sector because *AM* labor demand moves positively with productivity changes. Hence, the *AM* output is more responsive than *PCC* output to changes in k. Therefore, the smaller is Np, the more responsive is aggregate output to k. In general, given a fixed money stock, the price level will be more responsive to k if the output level is more responsive to k (in the opposite direction, of course). Hence, the smaller is N_P, the more correlated are k and p.

Firms are more willing to supply contracts when the $V(k|p)$ is small because the im-

plicit cost of these contracts in terms of lost information and flexibility is less when p is a good signal for k. Combining Proposition 2 and the Lemma, one can characterize the supply price, S, for contracts as follows.

PROPOSITION 3: *Under the assumptions listed in the Lemma,*

(*i*) *S is strictly increasing with N_p if $\sigma_e^2/\sigma_k^2 > 0$ and*

(*ii*) $S = Ez, \forall N_P$, *if* $\sigma_e^2/\sigma_k^2 = 0$.

Using Propositions 1 and 3, the graph of aggregate demand and supply of *PCC*s is given in Figure 1. The index q and the fraction of the economy that contracts (N_P) are measured along the horizontal axis; both variables are defined only upon the unit interval. The net income (in real units of consumption) awarded to unions under *PCC*s is measured on the vertical axis.

Each point along the demand curve $D(q)$ states the fraction of unions that desire to contract at a particular level of net income. The less risk averse prefer the *AM* alternative. The aggregate supply curve $S(N_P)$ is downward sloping because it incorporates the effect of the size of the *PCC* sector on the signaling power of p. A point along $S(N_P)$ states the size of the contracting sector that would cause $E\pi_p$ to equal $E\pi_A$ at a particular level of T. The vertical intercept $S(0)$ is below Ez if p is an imperfect signal for k.

When the contract price is T^*, the same fraction N_p^* of firms and unions supply and demand contracts, respectively. Since firms are identical and indifferent between the two arrangements at the contract price T^*, the particular firms which end up in the *AM* is arbitrary. However, only the marginal union is indifferent between labor market arrangements, and the less risk averse choose the *AM*.

Although the equilibrium in the figure depicts coexistence, if the costs of contracting are sufficiently large and $D(0) > S(0)$, only the *AM* would exist in equilibrium. On the other hand, coexistence is possible even without heterogeneity among unions. If unions were identical, then $D(q)$ would be horizontal but may intersect $S(N_P)$ at some N_P between 0 and 1.

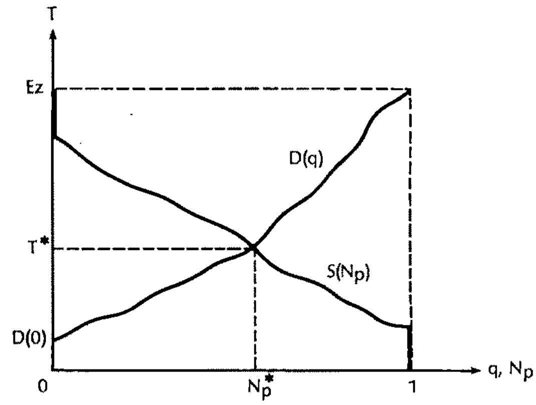


FIGURE 1

IV. Monetary Policy

Macroeconomic models with nominal wage contracts generally provide a channel for monetary policy to have real effects. For policy analysis, the current model has two advantages over models such as Fischer's. One, the agents' utility functions are explicit; therefore, optimal policies can be chosen to maximize a natural welfare criterion, Pareto optimality. Two, the long-run effects of alternative policy rules can be examined since the entire model can be expressed in terms of policy-invariant parameters.

Fischer assumes that contract duration exceeds the amount of time necessary to implement monetary policy. Under the additional assumption that the economy's stochastic structure is perceived as independent of monetary policy, Fischer derives the "optimal" money supply rule that minimizes the squared deviations of contractual output from the *AM* level. In part, this rule includes the complete offset of monetary shocks.

In the current model, the optimal monetary policy is straightforward.[5] The *AM* obviously produces the efficient, *ex post*, output level, but the contractual sector does not. The obvious policy objective should be to

[5]This model has the unusual property that positive monetary shocks reduce *PCC* output because high prices are interpreted as signaling low productivity inducing less employment. This feature is to be expected when productivity inferences are based upon the price level (see, for example, Canzoneri and Sibert).

induce the contractual sector to produce the efficient output level. This objective can be achieved by varying the money stock to assure that $E(k|p)=k$. The welfare of *PCC* agents would increase without affecting the welfare of *AM* participants. One way to achieve this objective would be to pursue Fischer's monetary rule. The complete offset of monetary shocks would make $V(k|p)=0$ and $E(k|p)=k$.

The effect of such a policy on the distribution of agents across sectors would be to increase the amount of contracting. If $V(k|p)=0$, the supply curve in Figure 1 becomes horizontal at $T=Ez$ and, in equilibrium, the *AM* disappears. This extreme conclusion can be tempered by the presence of firm-specific shocks, direct contracting costs, and the imperfect execution of monetary policy—all of which would reduce the demand for contracts. This increase in the size of the *PCC* sector induced by monetary policy parallels the increase in contract duration caused by monetary policy in the work of Gary Fethke and Andrew Policano, John McCallum, and myself (1983).

VI. Conclusions

In this paper, a general equilibrium model is developed in which firms and unions choose between *AM* participation and nominal contracting. Their choice depends upon the agents' attitudes toward risk, their production techniques, and their stochastic environment. There is an implied tradeoff between efficiency in the allocation of risk and production efficiency. However, in a broader sense, this may be interpreted as a tradeoff between the advantages of flexibility and precommitment.

The costs of contracting naturally increase with the size of the contracting sector because the signaling power of nominal magnitudes (for real variables) declines with their use for that purpose. Monetary policy that naively ignores its impact on the stochastic structure encourages contracting by making the expected costs of contracting smaller.

The model focuses on the coexistence of two labor market arrangements used for the production of a single good. One possible extension of the model would be to consider a two-good economy. Two different production processes $(f'\phi')$ would give rise to different levels of contracting in each sector. Different monetary rules would effect not only labor market arrangements, but also the composition of aggregate output. A more important but more difficult extension would be to increase the types of contracts available to labor market participants.

REFERENCES

Azaradis, Costas, "Escalator Clauses and the Allocation of Cyclical Risks," *Journal of Economic Theory*, June 1978, *18*, 199–255.

Barro, Robert J., "Long-Term Contracting, Sticky Prices, and Monetary Policy: A Comment," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

Cantor, Richard, "A Macroeconomic Model with Auction Markets and Nominal Contracts," Working Paper No. 84–2, Ohio State University, 1983.

Canzoneri, Matthew, "Labor Contracts and Monetary Policy," *Journal of Monetary Economics*, April 1980, *6*, 241–55.

______ **and Sibert, Anne,** "The Macroeconomic Implications of Labor Contracting with Asymmetric Information," International Finance Discussion Paper No. 248, Board of Governors, 1984.

Diamond, Peter and Stiglitz, Joseph, "Increases in Risk and Risk Aversion," *Journal of Economic Theory*, July 1974, *8*, 337–360.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Fethke, Gary and Policano, Andrew, "Long-Term Contracts and the Effectiveness of Demand and Supply Policies," *Journal of Money, Credit, and Banking*, November 1981, *13*, 439–53.

Gray, JoAnna, "On Indexation and Contract Length," *Journal of Political Economy*, February 1978, *86*, 1–18.

McCallum, John, "Stabilization Policy and Endogenous Wage Stickiness," *American Economic Review*, June 1983, *73*, 414–19.

Phelps, Edmund and Taylor, John, "Stabilizing Powers of Monetary Policy under Rational Expectations," *Journal of Political Economy*, February 1977, *85*, 163–90.

Predictable Behavior in Financial Markets: Some Evidence in Support of Heiner's Hypothesis

By FRED R. KAEN AND ROBERT E. ROSENMAN*

Recently, Ronald Heiner (1983) has proposed a theory of predictable behavior which has its roots in what Heiner labels a competence-difficulty gap (hereafter called a *C-D* gap). This gap is a measure of the spread between an economic agent's competence to make an optimizing decision and the difficulty of the decision problem. Heiner argues that as the *C-D* gap widens, the agent is increasingly likely to follow rule-governed behavior, and it is this rule-governed behavior that produces observed regularities in economic behavior.

If there is no *C-D* gap, Heiner argues that rule-governed behavior would disappear and, along with it, the observed regularities. Instead, *unpredictable* behavior, characterized by constant perturbation, would replace predictable behavior.

Furthermore, the dynamic properties of Heiner's theory are characterized, in his words, by sudden "switching" behavior. This switching behavior is described as a sudden reversal of previous behavior and a persistent movement in another direction or behavior of a different kind.

One of the casual or intuitive examples Heiner offers for support of his theory is the "switching between buying and selling strategies in financial markets, resulting in sudden movement in stock prices" (p. 582). This paper explores this idea and argues that there is more than intuitive support for Heiner's speculation; that, in fact, the behavior of daily price changes in financial markets conforms better to the prediction of Heiner's theory than to standard efficient market theory predictions.

*Professor of Finance, University of New Hampshire, Durham NH 08324, and Assistant Professor of Economics, Washington State University, Pullman, WA 99164, respectively. We are grateful to Ronald Heiner for extensive discussions and comments, and to an anonymous referee for helpful comments.

I. The Conventional View

Standard efficient market theory holds that all available information is reflected in the current price of a financial asset. Price changes other than those which are expected (trend) occur because news arrives which changes economic agents' expectations. If news arrives in a random manner, then price changes will also exhibit random behavior. (See Richard Brealey and Stewart Myers, 1984.)

Tests of efficient market theory are frequently made by testing the operational hypothesis that sequential price changes, adjusted for trend, are independent of one another. Lately, such tests have been described as tests of a speculative efficiency hypothesis rather than an efficient market hypothesis. This labeling change was made in order to emphasize that the absence of independence in sequential prices was not sufficient to conclude that markets were inefficient or irrational in the sense that it would be possible to earn risk-adjusted excess returns.

Early tests of the efficient market hypothesis (see Eugene Fama 1970, for an excellent summary) generally concluded that while "marginal" price change dependency was present, a weak-form (informational) market efficiency hypothesis could not be rejected if the operational hypothesis was recast into a fair-game (zero expected profits) hypothesis. Eventually, market efficiency became the conventional wisdom of financial economics because one study after another produced the same result: an inability to reject a fair-game efficiency hypothesis. Yet, an ever

present indication of "weak" statistical dependency in the price change time-series was always there. Perhaps the most nagging reminder of potential dependency was the continual ability of *ad hoc* trading rules to generate speculative profits (see Richard Stevenson and Robert Bear, 1970, futures markets; Sidney Alexander, 1961, common stock; Fama and Marshall Blume, 1966, stocks; G. Geoffrey Booth et al., 1981, foreign exchange rates; Booth and Kaen, 1979, gold and silver; and Stephen Figlewski and Thomas Urich, 1983, Treasury bill futures).

As empirical work which reported statistical dependencies in financial asset prices continued to appear, financial theorists responded with a variety of explanations for the empirically documented phenomenon. These explanations have usually modified the information assumptions of the original model.

Stanley Black (1976), for example, assumed that information did not arrive in continuous small units but, instead, in nonrandom large chunks. This assumption of Black produces a jump-like price change process and has been examined by, among others, George Oldfield et al. (1977). Although such a process can account for nonnormal price change distributions and discrete shifts in mean expected returns, the process is apparently unable to explain the persistent long-term dependence in security price changes which we later address in this paper (see Bruce Fielitz, 1979).

Information-based explanations that do account for statistical dependence in price change series have generally relaxed the assumption that information is disseminated at an infinite speed. A finite adjustment speed is traced to market imperfections such as transaction costs and institutional rigidities as in Avraham Beja and Nils Hakansson (1977) or to information asymmetries as in the informed vs. uninformed traders of Sanford Grossman and Joseph Stiglitz (1976) and Dale Morse (1980), and across securities, as in the work of John Groth (1979).

What we choose to relax is the assumption that all individuals have the same capacity to interpret an identical information set. By relaxing the assumption of homogeneous information-processing skills, we are able to explore the effects of agent-specific uncertainty in distinct contrast to market risk on security prices.

II. A *C-D* Gap Model of Switching Strategies in Financial Markets

Heiner's *C-D* gap is predicated on the idea that agents must make decisions for which the difficulty of the problem exceeds their competence. For those decisions, the decision process becomes one of making a decision under uncertainty rather than risk. It is the introduction of uncertainty into the decision-making process that, Heiner argues, leads to persistent price movements punctuated with sudden changes in directions.

Under Heiner's hypothesis, uncertainty, u, is a function of an agent's perceptual abilities, p, and the complexity of its environment, e. He argues that u is negatively related to p and positively related to e. One of the determinants of the decision environment is news, n. Therefore, with fixed perceptual abilities, $e = e(n)$; $e' < 0$.

We define news as previously unavailable information about the distribution of expected asset returns. Under this definition, news may either increase or decrease the risk-adjusted expected value of an asset. Consequently, we wish to argue that news decreases environmental complexity because it provides information for correctly revising previously held expectations.

Our definition of news is particularly appropriate for analyzing financial markets because participants in these markets establish and reestablish portfolios in response to the news they receive and their perceptual ability to interpret the news correctly. If we assume a nonhomogeneous distribution of perceptual abilities across agents, then the condition for agents to react differently to news is established. The agents or groups of agents will switch behavior (for example, buying to selling) at different times and this nonsynchronous switching will cause persistent price movements and, eventually, sudden direc-

tional changes. This can be put in the context of Heiner's model.

Using Heiner's notation, let

$$u = u(p, e(n)) \tag{1}$$

with $u_1 < 0$ and $u_2 > 0$. Furthermore, let $\Pi(e)$ be the probability that an agent's action (for example, to buy or hold an asset) is correct, and $(1-\Pi(e))$ be the probability that an agent's action would, in reality, be incorrect.

Whether the agent will actually buy or hold if it is called for, however, depends on the level of uncertainty facing the agent. Therefore, the probability that the agent will take an action when it should be taken is conditioned on this uncertainty and is represented as $r(u)$. The conditional probability of the agent buying or holding when such action is not called for is also conditioned on uncertainty and is represented as $w(u)$. It is expected that greater uncertainty lowers r and raises w, so that $r'(u) < 0$ and $w'(u) > 0$, which implies that r/w drops with greater uncertainty.

If the agent takes the action when it is the correct response, the agent will receive, on average, a gain represented as $g(e)$. If the agent moves when it is the incorrect response, the agent will sustain, on average, a loss represented as $l(e)$.

Within the above framework, an agent will buy (hold) whenever the expected gain from buying (holding) exceeds the expected loss. This decision is represented by

$$g(e)r(u)\Pi(e) > l(e)w(u)(1-\Pi(e)) \tag{2}$$

Rearranging (2) gives Heiner's Reliability Condition

$$R(p,e) = r(u(p,e))/w(u(p,e)) > \frac{l(e)(1-\Pi(e))}{g(e)\Pi(e)} = T(e). \tag{3}$$

Heiner calls the left-hand side the "reliability ratio." It can be interpreted as the actual reliability in responding to information about market values. The ratio measures the probability of correctly responding when doing so would produce a gain relative to the chance of mistakenly responding when doing so would produce a loss. In effect, it shows how an agent's uncertainty affects the relative probability of taking the correct action compared to the chance of taking an incorrect action. Heiner calls the right-hand side of (3) the "tolerance limit"; a minimum lower bound that is the unconditional expected loss relative to the unconditional expected gain. The value of $\Pi(e)$ determines the "riskiness" of the action. Agents benefit from acting only if (3) holds.

An alternative interpretation of $T(e)$ is as a "fairness ratio," which ranges from zero to infinity. If the market is "fair," then assets have an expected value of zero and $T(e)=1$. Values between 0 and 1 indicate a positive expected value based on the price and values of $T(e)$ greater than 1 would arise if an asset has a negative expected value. As the expected value becomes increasingly negative, for an agent to buy (or hold) an asset its reliability ratio must grow larger and larger. Thus, as an asset becomes less desirable in an expected value sense, an agent will buy or hold it only if his conditional confidence that it is the correct decision, its reliability ratio, grows large.

In a financial market context, $\Pi(e)$ may be interpreted as the probability that an asset is priced so as to provide a gain in net present value based upon its expected returns and systematic risk; $(1-\Pi(e))$ would be the probability that the asset is priced so as to provide a net loss. Then the risk-adjusted expected value of the asset is $g(e)\Pi(e) - l(e)\ (1-\Pi(e))$. For a properly priced asset, this sum should equal zero: the risk-adjusted net present value at the current price is zero (corresponding to $T(e)=1$ in condition (3) above). In a perfect (completely efficient) market, nonzero values would be immediately recognized and corrected without any mistakes being made. Price changes in this perfect market would adjust only to independent news flows, and thus would be independent of previous price changes (adjusted for *expected* changes) and would conform to a martingale process. Under

standard Bayesian decision theory, agents respond optimally to any information relevant to ascertaining an asset's risk-adjusted expected value. They, therefore, always act to maximize the risk-adjusted expected value conditioned on received news, however imperfect. Bayesian optimizers always respond correctly and instantly (in an expected value sense). However, once agents are allowed to respond imperfectly to (imperfect) information, an immediate response will not necessarily be beneficial, even when information is costlessly available. Traders may thus delay action.

Suppose, however, that at a given point in time an asset is not equilibrium priced; it has a positive (or negative) net present value. Depending on the values of $\Pi(e)$, $l(e)$, and $g(e)$; $T(e)$ will now take on values from 0 to ∞ and the conditions are met for behavioral change to take place. Recall that this behavior depends on the investor's reliability ratio which, in turn, depends on the perceptual ability of the agent and the information (news) that the agent has about whether $T(e)$ is greater or less than one.

The information available to the agent may be unavailable to others (inside information), or may be available to everyone (public). Our primary interest is with what the agent is able to do with publicly available information; it is usually conceded that inside information can be used to identify nonzero net present value investments.

What happens under our assumptions is that, for many traders, $R > T$ is violated initially as news about the true value of the asset first enters the market. More information increases R at the same time T adjusts towards unity (its equilibrium value). Thus, at some point, R surpasses T, although at different times for different individuals.

Because some agents are better able than other agents to interpret the information coming to the market place, these "superior" agents will exhibit, other things being equal, a larger reliability ratio. Consequently, these superior agents will "switch" positions prior to the "average" agent who is waiting for additional news. The early switching by the superior agents will be followed by later switching as more information comes to the market and, in the process, a persistent change in prices will be observed.[1,2]

Simultaneously with the arrival of additional news and the partial price adjustments towards a new equilibrium will be a change in $g(e)$, the possible gain and in $l(e)$, the possible loss. At the same time, $\Pi(e)$ adjusts and $T(e)$ moves towards unity. The possible gain and loss adjust because the price does. For example, with a stock, the biggest potential loss is the amount spent on the stock. The probability $\Pi(e)$ changes with additional news flow, as outcomes become more certain. What is observed is a crescendo of trading activity and, as $T(e)$ finally reaches unity, an abrupt return to relative calm. In statistical terms, nonperiodic dependence should be expected in a time-series of financial asset price changes.

As an example, suppose we have an asset that is undervalued. In a financial market context, this means the net present value of the asset based on its expected return and systematic risk is positive, which means that the tolerance limit or fairness ratio is less than unity.

Assume there is a queue of perceptual abilities for the N traders in the market, with $p_1 > p_2 > \dots > p_N$, so agent 1 has the best perceptual abilities (in this case). Suppose all news is equally (and instantaneously) disseminated through the market (unlike an entropy model) so each trader faces the same market complexity. At the same time, all the news relevant to the fact that the asset is undervalued need not be revealed at once. For example, the first news may be that a firm has a new product forthcoming. Second,

[1] Grossman and Stiglitz have explored reasons why investors without information do not imitate investors with information. Their analyses with respect to the *imitation* question may be incorporated into our model by substituting low *C-D* gap investors for informed investors and high *C-D* gap investors for uninformed investors.

[2] The key here is *not* when or which agents receive more news (since all news is assumed publically available to all traders at the same moment), but the ability of agents to correctly respond to incoming news.

the characteristics of the product are leaked. Third, the product is revealed; and finally, the public's acceptance of the product is seen. Highly perceptive traders could recognize that this firm's stock is underpriced at the first or second information flows, while less perceptive traders may not be willing to act on this possibility until more (confirming) information is received. (For example, they may not react until the product is actually revealed or received.) By the assumptions of the model, $u^1 < u^2 < \dots < u^N$, where the superscripts indicate the uncertainty suffered by the ith trader.

In this example, $r(u)$ is the probability that the asset is purchased when it should be,[3] and $w(u)$ is the probability of still purchasing the asset when it should not be. Recall that the chances of correctly and mistakenly responding to incoming news are, respectively, inversely and positively related to uncertainty (i.e., $r'(u) < 0, w'(u) > 0$). Thus, $r(u^1) > r(u^2) > \dots > r(u^N)$ and $w(u^1) < w(u^2) < \dots < w(u^N)$. These together imply $R(p_1, e) > R(p_2, e) > \dots > R(p_N, e)$.

The relationship between agents' reliability ratios and the (market determined) tolerance limit gives the path of price movements. There are three possibilities. First, it could be that the asset is so obviously undervalued that for *all* traders, $R(p,e) > T(e)$. In this case, demand would push the price upward immediately, eliminating profit potentials. This is similar to the situation that would arise when no traders suffer a *C-D* gap. Then, for this example, $r(u) = 1$ and $w(u) = 0$ for all traders, and prices adjust immediately to equilibrium values. (Since $R(u) = \infty$ for all traders, they can all benefit from immediately reacting to incoming news regardless of how large $T(e)$ might be.)

A polar case is when news is so nebulous that even the trader with the best perceptual ability has a high level of uncertainty, so that $T(e)$ still exceeds $R(p_1, e)$. Since this implies $R(p,e) < T(e)$ for all traders, no one demands the asset and the price does not adjust. Only when additional news becomes available do prices move.

It is the intermediate case that is of most interest; when some traders reliably perceive that the asset is underpriced, and others do not.[4] Without loss of generality, assume $R(p_1, e) > T(e)$ and $R(p_i, e) < T(e)$ for $i = 2, \dots N$. Then, the first agent will demand the undervalued asset, initiating some movement. As more news flows to the market, more agents respond (because the extra news raises successively more $R(p_i, e)$ over $T(e)$), thereby increasing market demand as their reactions accumulate. Part of the news flow, could, of course be that other traders are buying the asset.[5] Increased purchasing forces the price up, affecting the potential gain or loss and the relative probabilities of the risk-adjusted gain or loss.

For the underpriced asset, at its current price the expected risk-adjusted present value exceeds zero, that is,

$$(4) \quad g(e)\Pi(e) - l(e)(1 - \Pi(e)) > 0.$$

The potential gain and loss and probabilities of each depend on the price. As the price increases, the left-hand side of (4) drops towards zero. ($T(e)$ converges towards unity from below.) Since there is *no guarantee* of a gain, we maintain that $0 < \Pi(e) < 1$. As these changes proceed, $g(e)$ probably decreases and $l(e)$ probably increases in magnitude, but is bounded by the price of the asset.

At the same time that prices partially adjust to incoming news, the reliability ratio $R(p_i, e)$ increases for the ith individual. More news implies less uncertainty. As news increases, it becomes clearer that the asset is underpriced. Consequently, agents are more likely to take the correct action, and less likely to mistakenly respond, so that $r(u)/w(u)$ grows.

[3] Buying the asset is the correct action because in an expected value sense the asset is undervalued. It does not guarantee, *ex post*, that purchasers make money.

[4] Most likely, perceptual queues will vary over time and by news-type and source. Thus it would not be possible for the market to establish a leader whom all traders follow. At the same time specialists should arise —those who concentrate in that type of asset where a small *C-D* gap gives them a comparative advantage.

[5] See again fn. 4.

Both the left-hand and right-hand side of (3) grow. The right-hand side approaches unity (the equilibrium value) from below while the left-hand side grows unbounded. More and more traders recognize the asset is underpriced. As successively more agents' reliability ratios pass the tolerance limit, there are repeated partial adjustments that collectively generate a persistent movement of the asset's price towards its new equilibrium value.

Analagously, if an asset is overpriced, its risk-adjusted expected value would be negative and the correct action for holders of the asset would be to sell. In an analysis similar to the underpriced asset, there would be persistent partial price adjustments downward until its risk-adjusted expected value is zero. Initially $T(e) > 1$, but it moves towards unity from above as the price adjusts. In either case, we should see an increasing volume of trading, reaching its peak at the mode of the perceptual abilities distribution, followed by a slowing and a period of relative inactivity at the equilibrium price. Trading could continue, but at zero expected value.

Clearly, the absence of *C-D* gaps is a limiting case of this model. With no difficulty in understanding news, $R(p, e) = \infty$ and all agents would immediately assimilate news and recognize an under- (or over-) priced asset. Adjustment to the equilibrium price (and equality in equation (4)) would be (nearly) instantaneous, and price movements would mimic the arrival pattern of news—most likely a random walk or jump process. Trading would continue, but the price would have already adjusted, eliminating positive risk-adjusted expected values. But, with *C-D* gaps, price adjustments are partial, even for public news made costlessly available to all traders at the same time. However, they continue persistently towards the equilibrium price until risk-adjusted expected value is zero.

In sum, the existence of *C-D* gaps in financial markets would cause persistent, nonperiodic movements or cycles in prices. Presented in the next two sections are a statistical tool for identifying nonperiodic cycles and some evidence of their existence in financial markets.

III. A Statistical Perspective

In the field of hydrology, there is a statistical procedure called rescaled range (R/S) analysis. The key technical references for this procedure are H. E. Hurst (1951), B. B. Mandelbrot (1972), Mandelbrot and J. R. Wallis (1969), and Wallis and N. C. Matalas (1970). The special attribute of R/S analysis is that it can identify *nonperiodic* cycles in time-series data. The statistical phenomenon of nonperiodic cycles is called persistent dependence; and it has been dubbed "memory" by many who use the technique.

A time-series that exhibits persistent dependence has not been generated by a white noise process. More importantly, a time-series that exhibits positive persistent dependence exhibits a tendency to move away from its mean value for an "extended" period of time before it changes direction and moves in that direction for another extended period of time.

The statistical measure of persistent dependence is called the Hurst coefficient. The Hurst coefficient or statistic relates the ratio of the range (R) of cumulative departures from a trend line of a time-ordered subsample of a time-series and the standard deviation (S) of that subsample time-series to the product of a constant and the size (n) of the subsample time series. This relationship is set forth in

$$(5) \qquad R/S = R(t,n)/S(t,n) \sim cn^h,$$

where $R(t,n)$ = the range or difference between maximum and minimum values within a subsample time period (t) of n consecutive observations of cumulative departures from a trend line; $S(t,n)$ = standard deviation of the subsample time-series (t) where n indicates the number of observations in t; c = constant; n = number of consecutive observations or length of the subsample time period; and h = Hurst coefficient.

Equation (5) may be readily transformed into the regression equation (6) for empirical estimating purposes. (See the Appendix for a technical presentation of this material.)

$$(6) \qquad ln(R/S) = ln(c) + hln(n)$$

For a stationary time-series, the Hurst coefficient may take on values between 0 and 1. If it is 0, then the time-series is a perfect sine wave. If it is greater than 0 but less than .5, then the time-series has irregular short cycles.

For coefficient values greater than .5 but less than 1, the time-series is characterized by long irregular cycles and is said to have persistent dependence. If the value of the coefficient is .5, then the series is a classic white noise series.

TABLE 1—HURST COEFFICIENTS FOR 200 COMMON STOCK RETURN SERIES

Frequency Distribution of Hurst Coefficients	
0.401–0.450:	1
0.451–0.500:	35
0.501–0.550:	93
0.551–0.600:	63
0.601–0.650:	7
0.651–0.700:	1
Summary Measures	
Range:	0.448–0.651
Average R^2:	9.99+
Row R^2	0.98

Source: Greene and Fielitz (p. 345).
[a]GH 10 procedures; 1220 daily returns.

IV. "...[E]mpirical results in search of a rigorous theory"[6]

An early application of R/S analysis to financial data was made by Myron Greene and Fielitz (1977). They applied R/S analysis to the daily return series for the first 200 common stocks listed on the New York Stock Exchange. Their sample period ran from December 23, 1963 to November 29, 1968.

Table 1 summarizes the results reported by Greene and Fielitz. Of the 200 common stocks they examined, 164 had Hurst coefficients greater than .5; therefore, they found evidence of long-term persistent dependence.

The next application of R/S analysis was by Booth and Kaen (1979) to spot gold prices. They reported that the Hurst coefficient for a time-series of daily returns on gold prices was .642 for the period 1969 through 1980. Once again persistent movements in a single direction (nonperiodic cyclic behavior) were uncovered.

Booth et al. (1982) extended their exploration for persistence to the foreign exchange market. In this study, which used daily changes in spot exchange rates for the period July 1, 1973 to June 30, 1979, they reported Hurst coefficients of .67, .57 and .55 for the U.S. dollar price of, respectively, British pounds, French francs, and West German marks. The results provided strong support for the persistent dependency hypothesis.

In addition to the stock market, the foreign exchange market, and the gold market, persistent dependence has also been reported in the futures markets. Billy Helms et al. (1984) applied R/S analysis to both inter- and intraday price changes for soybean, soybean oil, and soybean meal futures contracts. Hurst coefficients reported ranged in value from .558 to .711; higher values were found for the interday changes than for the intraday changes.

All of the R/S studies reported reject a speculative efficiency hypothesis of sequential price change independence. The studies do not, however, reject a fair game version of the efficient market hypothesis. What the studies do raise are questions about the underlying economic process which is generating the observed behavior. Heiner's theory supplies one possible explanation.

V. Statistical Persistence and Predictable Behavior

Heiner's *C-D* gap provides an explanation for the observed statistical persistence and switching behavior in financial markets without implying the existence of a "money machine," asymmetric information, a finite speed of information dissemination, or any costs of receiving information. Highly perceptive individuals react to news early, the less perceptive wait for more news. The flow of movement causes a *series* of partial adjustments towards the new equilibrium asset price. This introduces persistence in price movements. However, because there is no indication as to the length or magnitude of the adjustment period, the theory does not

[6]Fama (p. 389).

allow for a money machine. It is this implied existence of a money machine which has plagued past explanations of "memory" in financial market price data.

By explicitly introducing uncertainty and using it to predict logical rule governed behavior, Heiner has added an element heretofore missing in "fair game" market efficiency theory. The postulated *C-D* gap not only explains the presence of nonperiodic regularities in price changes, but also elucidates the always recognized possibility for superior fundamental analysts to consistently outperform the market. These analysts simply have a relatively smaller *C-D* gap!

APPENDIX

The symbol R/S stands for rescaled range, $R(t,n)/S(t,n)$. Its components may be described as follows. Let $X(t)$ be a discrete, stationary time-series for which there are T observations. Denote $X^*(t)$ as the series' cumulative sum such that:

$$\text{(A1)}\quad X^*(t)=\sum_{u=1}^{t} X(u) \quad \text{for} \quad 1\leq t\leq T.$$

Then,

$$\text{(A2)}\quad R(t,n)=\max_{0\leq u\leq n}\left[X^*(t+u)-(X^*(t)+(u/n)(X^*(t+N)-X^*(t)))\right] - \min_{0\leq u\leq n}\left[X^*(t+u)-(X^*(t)+(u/n)(X^*(t+N)-X^*(t)))\right],$$

$$\text{(A3)}\quad S(t,n)=\left[(1/n)\sum_{u=1}^{n}X^2(t+u) - (1/n^2)\left(\sum_{u=1}^{n}X(t+u)\right)^2\right]^{0.5},$$

where t is any starting point and n is any subseries (sample) such that $n\leq(T-t)$. The $R(t,s)$ is defined to be the subseries sequential range and reflects adjustments for the removal of the subseries trend. The standard deviation of the subseries that is defined by t and n is $S(t,n)$. We obtain $R(t,n)$ from the cumulative series X^*, but $S(t,n)$ comes from the original series (X).

Mandelbrot and Wallis suggest that R/S is asymptotically related to n, the subseries size. Specifically,

$$\text{(A4)}\quad R/S=R(t,n)/S(t,n)\sim cn^h,$$

where c is a constant and h is the Hurst coefficient.

REFERENCES

Alexander, Sidney S., "Price Movements in Speculative Markets: Trends or Random Walks," *Industrial Management Review*, May 1961, *2*, 7–26.

Beja, Avraham and Hakansson, Nils H., "Dynamic Market Processes and Rewards to Up-to-Date Information," *Journal of Finance*, May 1977, *32*, 291–303.

Black, Stanley, "Rational Response to Shocks in a Dynamic Model of Capital Asset Prices," *American Economic Review*, December 1976, *66*, 767–79.

Booth, G. Geoffrey and Kaen, Fred R., "Gold and Silver Spot Prices and Market Information Efficiency," *Financial Review*, Spring 1979, *14*, 21–26.

______, **Kaen, Fred R. and Koveos, Peter F.**, "Foreign Exchange Market Behavior: 1975–1978," *Rivista Internazionale Di Scienze Economiche E Commercial*, April 1981, *28*, 311–26.

______, ______, **and** ______, "*R/S* Analysis of Foreign Exchange Rates Under Two International Monetary Regimes," *Journal of Monetary Economics*, December 1982, *10*, 407–15.

Brealey, Richard and Myers, Stewart, *Principles of Corporate Finance*, New York: McGraw-Hill, 1984.

Fama, Eugene F., "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, May 1970, *25*, 383–417.

______ **and Blume, Marshall**, "Filter Rules and Stock Market Trading Profits," *Journal of

Business, January 1966, *39*, 226–41.

Fielitz, Bruce, "Empirical Research on Capital Markets: Discussion," *Journal of Finance*, May 1979, *34*, 465–69.

Figlewski, Stephen and Urich, Thomas, "Optimal Aggregation of Money Supply Forecasts: Accuracy, Profitability, and Market Efficiency," *Journal of Finance*, June 1983, *38*, 695–710.

Greene, Myron T. and Fielitz, Bruce, "Long-Term Dependence in Common Stock Returns," *Journal of Financial Economics*, May 1977, *4*, 339–49.

Grossman, Sanford S. and Stiglitz, Joseph E., "Information and Competitive Price Systems," *American Economic Review Proceedings*, May 1976, *66*, 246–53.

Groth, John C., "Security-Relative Information Market Efficiency: Some Empirical Evidence," *Journal of Financial and Quantitative Analysis*, September 1979, *14*, 573–93.

Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, *73*, 560–95.

Helms, Billy P., Kaen, Fred R. and Rosenman, Robert E., "Memory in Commodity Futures Contracts," *Journal of Futures Markets*, Winter 1984, *4*, 559–67.

Hurst, H. E., "Long-Term Storage Capacity of Reservoirs," *Transactions of the American Society of Civil Engineers*, No. 1, 1951, *116*, 770–99.

Mandelbrot, B. B., "Statistical Methodology for Nonperiodic Cycles: From Covariance to R/S Analysis," *Annals of Economic and Social Measurement*, July 1972, *1*, 259–90.

______ **and Wallis, J. R.,** "Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long-Run Statistical Dependence," *Water Resources Research*, October 1969, *5*, 967–88.

Morse, Dale, "Asymmetrical Information in Securities Markets and Trading Volume," *Journal of Financial and Quantitative Analysis*, December 1980, *15*, 1129–48.

Oldfield, George S., Rogalski, Richard J. and Jarrow, Robert A., "On Autoregressive Returns," *Journal of Financial Economics*, December 1977, *5*, 389–418.

Stevenson, Richard A. and Bear, Robert M., "Commodity Futures: Trends or Random Walks?," *Journal of Finance*, March 1970, *25*, 65–81.

Wallis, J. R. and Matalas, N. C., "Small Sample Properties of H and K Estimators of the Hurst Coefficient h," *Water Resources Research*, December 1970, *6*, 1583–94.

The Choice Among Medical Insurance Plans

By YAEL BENJAMINI AND YOAV BENJAMINI*

Theoretical studies dealing with the efficiency of medical insurance have appeared in the economic literature since Kenneth Arrow's seminal article (1963), yet few deal with the choice among plans. This is because until recently most people in the United States were offered only one plan as a job-related fringe benefit, which meant that choosing another plan implied substantial financial loss. Empirical study concerning the effect of socioeconomic factors on the choice of plan was thus almost impossible.

Following the Health Maintenance Organizations (HMO) Act of 1973, many employers now offer their workers group policies for several plans, with no tax advantage to one method or another. The numerous different plans can be divided into two major categories: conventional insurance (*CI*) schemes and HMO plans. The *CI* schemes are characterized by fee-per-service, third-party reimbursement of most medical expenses made on behalf of the insured (for example, the Blue Cross/Blue Shield/Major Medical plans).[1] Under a HMO plan, the insured has access to a comprehensive range of health services in time of need in return for per capita periodic payments.

This paper analyzes the choice between the two types of insurance. An interpretation of Richard Zeckhauser (1970) implies that for a homogeneous group, an optimal HMO plan is Pareto superior to the optimal conventional one. It is shown here that the opposite may be true for a nonhomogeneous group, depending on how heterogeneous the group is. Although for members of a homogeneous group a HMO plan is welfare superior, members of a heterogeneous group may be better off with a *CI* scheme.[2] This discrepancy between the results for homogeneous and heterogeneous groups has theoretical importance. Our study suggests that where people can choose between the two types of plans, the group that is formed in a specific HMO plan is more homogeneous than its counterpart in a *CI* scheme. We conclude by discussing the practical consequences of these findings for the design of medical insurance plans.

I. *CI* vs. HMO for Homogeneous Groups

Here we compare the two types of medical insurance for a large group of identical individuals.[3] Following Zeckhauser's case I and Arrow (1976), we assume that a representative individual purchases a policy for an actuarially fair premium, and in time of need decides how much medical services to consume. These services are purchased on a fee-per-service basis, but the price the insured actually pays per unit (the coinsurance rate) is lower than the real market price. Increasing the co-insurance rate will result, simultaneously, in a welfare gain, as the medical care subsidy that causes moral hazard (see Arrow, 1976) is reduced, and a welfare loss because of increasing risks.

*Department of Economics and Department of Statistics, respectively, Tel Aviv University, Ramat-Aviv 69 978, Tel Aviv, Israel. This research was partially supported by funds granted to the Foerder Institute for Economic Research by Bank Leumi Le-Israel. The work includes part of the first author's Ph.D. thesis done at Princeton University under the guidance of Edwin Mills to whom thanks are due. We also thank Gary Truhlar of the University of Pennsylvania for valuable help in organizing the data for the analysis.

[1]All the *CI* systems are variations on this theme with different premiums, coverage limit, deductibles, or co-insurance schemes.

[2]The analysis for a homogeneous group is equivalent to using a representative individual, a widespread practice in theoretical studies (see, for example, Zeckhauser and Arrow, 1976). In view of the results of this work, using a representative individual might have misleading implications.

[3]Large in the sense that the law of large numbers is applicable; identical in the sense that they have the same income, distribution on states of health, and utility function.

This second-best type of plan can be compared to an optimal plan where the expected utility of any individual is maximized under a collective budget constraint (which is also the expected budget constraint for each member). Such a plan is described in Zeckhauser's case II, where optimality is achieved in a perfect contingent claims market. The individual gets a lump sum state-dependent insurance payment and purchases the medical service at the market price. Thus overuse (moral hazard) due to reduced price is prevented, while risk spreading is optimal. As Zeckhauser puts it, this plan is "the best that can be done with any actuarially fair plan" (p. 20). However, it is not materialized because it "requires the insurance plan to be able to distinguish unambiguously among all medical conditions... encourage its insureds to evade through misclassification, lead to errors in classification that would create injustices as well as inefficiencies" (p. 20).

While recognizing the problematic nature of such a contingent-claim market insurance plan, we suggest here that it is still possible to achieve optimality by employing a different organizational structure; in case of illness, the insured gets the needed medical treatment in-kind, instead of getting the money to purchase it. This is the essence of HMO plans where a contract is made between a group of insureds and a team of medical practitioners; for a periodic per capita actuarially fair payment, the patient receives treatment free of charge in time of need. The services given are the members' optimal choice *ex ante*, but dictated to them *ex post*, so there is no way the patient can consume more and create moral hazard. Thus, consumption of medical care and all other goods as well as risk distribution is optimal, implying that the HMO plan represents a Pareto optimal resource allocation.

Comparing the *CI* and the HMO plans, it is clear that the latter affords a higher level of welfare; in the *CI* plan the disposable income for medical and all other goods and services is the same for all states of health, while for the HMO plan's insureds, it can vary. This means that the levels of consumption of medical services and all other goods in the best *CI* plan are feasible but not optimal for the HMO plans' insureds.

II. *CI* vs. HMO for Heterogeneous Groups

Medical insurance is purchased primarily through job-connected group policies, and the results derived in Section I assume that such groups are homogeneous. However, we show below that the optimality of the HMO schemes for a group policy relies crucially on this assumption, thus questioning the validity of its usage in our problem. A HMO plan is very rigid: all members pay the same premium and receive the same services when sickness occurs. If income, tastes, or any other factory that affects the demand differs, this is inefficient because consumption may be above or below the preferred level. The *CI* plan is more flexible. Although the terms of insurance are identical for all, patterns of consumption may differ, giving the *CI* system an efficiency advantage.

This is demonstrated graphically in Figure 1, where the *DD* curve represents the demand for medical services in a specific state of health. If the market price is 1 and *SM* is a perfectly elastic supply curve, the optimal quantity is v_s^*. This is also the quantity supplied by the HMO plan; the HMO policy holders observe the totally inelastic supply curve S_p.

Figure 1 also describes the parallel situation for *CI* policy holders. These patients, with the same demand curve *DD*, encounter a perfectly elastic supply curve at price q, which is the co-insurance rate adopted by this *CI* plan. The quantity demanded is $\hat{v}_s$, where $\hat{v}_s > v_s^*$. The darkened triangle is the welfare loss introduced by the wedge between the producers' price and the consumers' price. The magnitude of the welfare loss depends on the size of $1-q$ and the elasticity of demand for medical services at this state of health. The total welfare loss for a *CI* patient is the expected sum of these losses for all states of health. As there is no deadweight loss associated with the HMO plan, it is clear that for one individual, or for a homogeneous group of such individuals, the HMO plan is welfare superior to the *CI* plan.

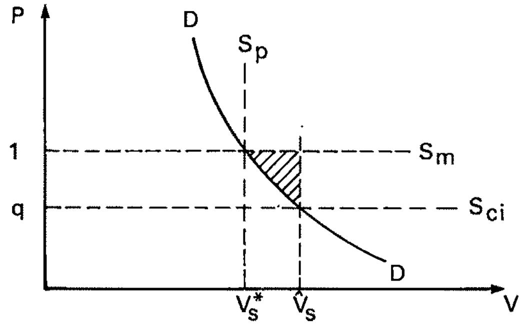


FIGURE 1. WELFARE LOSS WITH THE *CI* PLAN

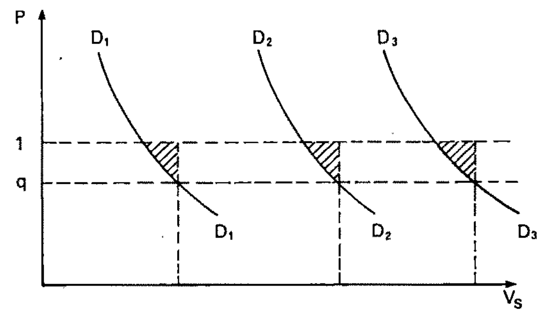


FIGURE 2. WELFARE LOSS FOR HETEROGENEOUS GROUPS IN THE *CI* PLAN

Now, consider a heterogeneous group consisting of three equal-sized subgroups, with DD_i representing the demand curve of the ith group for medical services at the same state of health. In Figures 2 and 3a, b, DD_1 represents the subgroup with the lowest income and DD_3 the subgroup with the highest income.[4] As DD_2 represents the demand curve for the median insured, this is the relevant curve for the insurance plan.[5]

Figure 2 shows the welfare loss for these groups in a *CI* plan with a co-insurance rate (and premium) that fits the median insured. It is apparent that the size of the deadweight loss does not depend on the distance between the groups; the sum of losses is not significantly different if the demand curves are DD_1' and DD_3' (as shown in Figure 3b), instead of DD_1 and DD_3. (In cases where the demand functions of the different groups are just shifts, the losses are unaffected by the actual distance.) In the HMO scheme, the situation is very different. With the same three subgroups, there is a welfare loss associated with consumption above or below the most desired level of medical services, shown in Figures 3a and 3b by the darkened areas. It is obvious that the further the demand curves are from each other (i.e., the more heterogeneous the group), the larger is the corresponding aggregate welfare loss, with no such loss to the median insured.

We can summarize the graphic comparison between the two plans as follows. Al-

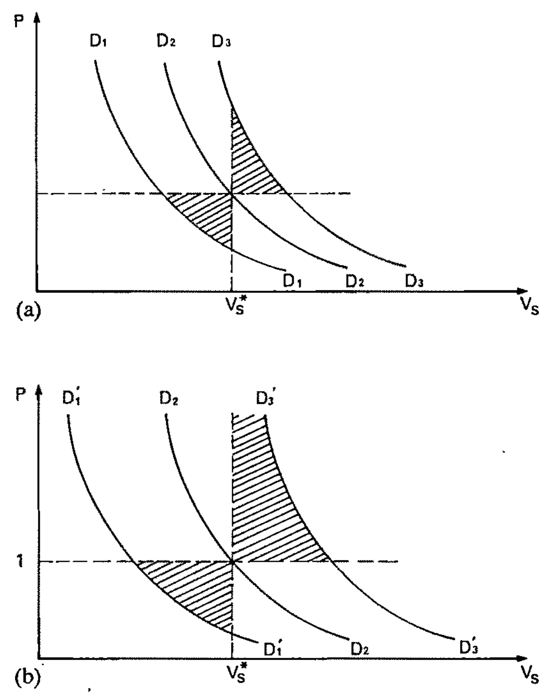


FIGURE 3. WELFARE LOSS FOR HETEROGENEOUS GROUPS IN THE HMO PLAN

[4]Equivalently, DD_1 may represent the group with low preference to medical care, and DD_3 the group with high preference, either due to tastes or to subtle physical differences.

[5]In job-connected group policies, the terms of the policy are usually affected only by marital status or family size. Using the median insured can be justified on the grounds that when an insurance firm tries to attract groups and compete with other plans to gain the favor of the majority of workers, it must appeal to the median worker.

though a homogeneous group is definitely better off in an HMO plan, a heterogeneous group may be worse off. The advantage of the HMO plan declines as differences in tastes, income, or risk prospects within the group increase. When heterogeneity is sufficiently strong, a *CI* plan may be better. Any specific individual, however, may determine whether an HMO scheme is better for him than a *CI* scheme by ascertaining whether any of the available HMO plans is close enough to his preferred plan.

III. The Empirical Evidence

Data on the 6168 employees of the University of Pennsylvania are used to explore empirically the differences in homogeneity among groups of insureds. The university's employees are offered a choice of six health insurance schemes, of which we compare here the members of two: a *CI* scheme—Blue Cross/Blue Shield/Major Medical, and a HMO scheme—The Philadelphia Health Plan. The employer's contribution to the medical insurance premium is independent of the chosen plan. The health scheme enrollment can be changed every six months without any reduction in coverage. Finally, each person associates daily with people who choose other plans and has access to knowledge about the pros and cons through open houses and written material. Thus we may assume that most of the employees prefer the plan they are enrolled in.[6]

We compare the homogeneity of the two groups by comparing the distributions of income, age, and the level of education. This is done by comparison of measures of the absolute and relative variation (spread) of the two groups. Along more traditional measures such as the variance, the variance of the logarithm of the variables, and the coefficient of variation, we use the interquartile range of the variables and of the logarithm of the variables. The latter are more robust against extreme values, which might be of doubtful relevance but nevertheless exist in the data at hand. The distributions of income and age for the two groups are visually compared in Figure 4 using boxplots.[7] Data about education are given in terms of eight partially ordered categories that were ordered and ranked. We thus supplement the analysis of education using the ranks by comparing the entropy of the two groups. Entropy is a measure of homogeneity defined on the distribution of the population in the n categories:

$$\mathbf{P} = (P_1, \dots, P_n)$$

$$\text{by} \quad H(\mathbf{P}_n) = -\sum_{i=1}^{n} P_i \log P_i.$$

If all the population is concentrated in one category, all p_i's are 0 except for one which is 1, and $H(\mathbf{P}_n) = 0$. If the population is heterogeneous and uniformly divided in the categories $H(\mathbf{P}_n) = \log n$. Thus higher entropy indicates more heterogeneity. See Henri Theil (1971, p. 71) for details and interpretation.

The values of the various measures of homogeneity as computed for the two groups are given and compared in Table 1. The absolute measures indicate that the heterogeneity of the *CI* group is 30–50 percent larger than that of the HMO group. The difference indicated by the relative measures —those eliminating differences associated with the different levels of the variables—is 15–30 percent. (The latter comparison is especially important for age because the two groups have different age averages.) En-

[6] There is a possible bias because of spatial considerations when benefits are given in-kind. Since members of HMO plans have to travel to specific locations to receive the medical services, those living farther away might prefer a *CI* plan merely of the above consideration, which is not captured in the theoretical model. We thus limit the group of *CI* plan members used in the comparison to those living in neighborhoods (identified by Postal Zip Code information) from which the HMO plan attracts its members.

[7] A boxplot is a schematic display of the distribution. The top and the bottom of the box are the quartiles and the cut is the median. The lines connect the furthest points which are less than 1.5 times the interquartile range away from the quartiles. Outside points are displayed individually. For detailed description, see J. W. Tukey (1977).

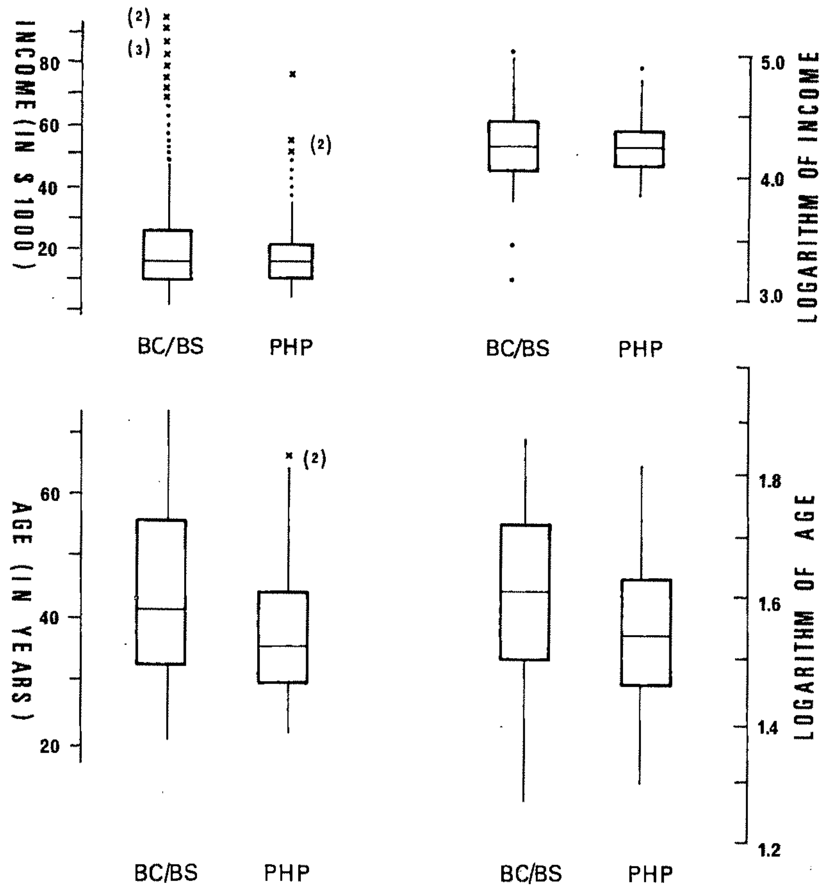


FIGURE 4. BOXPLOTS OF INCOME AND AGE (AND LOGARITHMS)

TABLE 1—COMPARISONS OF CONVENTIONAL INSURANCE (*CI*) PLANS AND HEALTH MAINTENANCE ORGANIZATION (HMO) PLANS

Variable	Type	Number in Group	Mean	Median	Standard Deviation	Ratio of Standard Deviations[a]	Inter-quartile Range	Ratio of Interquartile Ranges[b]	Coefficient of Variation
Income	*CI*	3279	20,874	17,400	12,628	1.38[c]	15,144	1.34	.366
(in Dollars)	HMO	491	19,141	17,400	9,134		11,300		.228
Log of	*CI*	3279	4.25	4.24	.233	1.23	.368	1.24	.003
Income	HMO	491	4.24	4.24	.190		.288		.002
Age	*CI*	3279	41.5	40	12.5	1.2	21	1.15	.091
(in years)	HMO	491	37.1	35	10.4		14		.079
Log of Age	*CI*	3279	1.60	1.60	.133	1.16	.225	1.3	.007
	HMO	491	1.55	1.54	.115		.171		.006
Education	*CI*	2754	4.87	5	2.47	1.04[c]	6	1.5	.244
(in ranks)	HMO	398	5.47	5	2.35		4		.221

[a]Significance of Ratio $\neq 1$ judged through F-test for equality of variances done under the assumption of normal populations.
[b]Significance of Ratio $\neq 1$ judged using the variances of the assymptotic distributions of the Interquartile Ranges (see David). For education binominal confidence limits were used.
[c]Inappropriate to judge significance because assumptions are clearly not met.

tropy for education is 1.97 (.01) for the *CI* group, and 1.71 (.04) for the HMO group (with asymptotic standard errors in parentheses).

The statistical significance of the above results is judged by various methods. Variances of the normal asymptotic distributions of the interquartile ranges (H. A. David, 1970) and of the entropy measures (Theil) are estimated. These variances are used to test whether the differences in these heterogeneity measures between the two groups are significant. For the variance measures, the F-test (based on the assumption that the distribution of the variables are normal) is used. For the analysis of education expressed in ranks, where the asymptotic distribution of the interquartile range is nonnormal, confidence bounds are computed for the interquartile ranges of each group, and the two were checked for overlap. The conclusions of these tests were the same. The differences described previously were all significant at least at the .0001 level.

We therefore conclude that for the two groups at hand, all the comparisons of the measures of homogeneity of income, age, and education, show that the *CI* group is more heterogeneous than the HMO group.

V. Concluding Remarks

The theoretical framework presented in this work suggests that an HMO plan is a more efficient way of providing medical insurance for a relatively homogeneous group. Different groups may have different plans that best fit their members. People who do not find an HMO plan close enough to their ideal (or live far from the HMO medical center) would be better off using the *CI* plan. In the set of data under research it has been shown that where employees can choose among several plans they cluster in relatively homogeneous groups in a specific HMO plan. This finding can be explained by the theory above that the members perceive the importance of homogeneity in HMO plans. A competing explanation argues that newcomers to the university are more likely to choose the less conservative HMO scheme, forming a group that is more homogeneous. The full explanation is probably a combination of those two (and possibly others), but in view of the additional work on this data (see our 1984 paper), we value the latter as less important.[8]

Using this theory to characterize the desired market for medical insurance, it is apparent that a uniform HMO plan of insurance cannot be optimal, as it is impossible for one plan to approximate the choice of most of the people when the group is as big as diversified as the whole nation. It implies that any governmental system that imposes a unique HMO-like structure on the entire society (such as the British National Health Services) is inefficient. Thus, the optimal system would consist of various HMO plans, coexisting with fee-for-service providers whose patients are covered by conventional policies. We recommend that, wherever spatially feasible, people should be encouraged to make an intelligent choice between as many available plans as feasible to enable more people to take advantage of their preferred HMO plan.

[8]In our earlier paper, we used a logit model to explain the choice of the plan. Using that model, a decrease of five years in the age (the difference of the two groups) is estimated to increase the probability of choosing an HMO plan by .007. Though the change was found to be significantly different than zero, it is too small to entirely explain the empirical findings.

REFERENCES

Arrow, Kenneth J. "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, December 1963, *53*, 941–73.

______, "Welfare Analysis of Changes in Health Co-Insurance Rates," in R. Rosett, ed., *Role of Health Insurance in the Health Services Sector*, Universities–NBER Conference Series, No. 27, New York 1976.

Benjamini, Y. and Benjamini, Y., "A Logit Model for the Choice Among Medical Insurance

Plans," Working Paper No. 31–84, Foerder Institute for Economic Research, Tel Aviv University, 1984.

David H. A., *Order Statistics*, New York: Wiley & Sons, 1970.

Theil, Henri, *Principles of Econometrics*, New York: Wiley & Sons, 1971.

Tukey, J. W., *Exploratory Data Analysis*, Reading: Addison Wesley, 1977.

Zeckhauser, Richard, "Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, March 1970, *2*, 10–26.

Is Statistical Discrimination Efficient?

By STEWART SCHWAB*

Neoclassical economists have advanced two general types of labor market discrimination models:[1] taste discrimination models and statistical discrimination models. Taste models include in the utility functions of employers, fellow workers or customers a desire to avoid members of certain groups. (See Gary Becker, 1957.) Under such an approach, discrimination cannot be characterized as either efficient or inefficient. One can compare the distribution of income and utility with an economy where people do not have a taste for discrimination. But until one decides the moral question of whether furthering the taste is acceptable, one cannot begin to ask whether society's resources are being placed in their most productive uses.

Statistical discrimination, by contrast, can affect efficiency. Statistical discrimination differs from the classic taste-for-discrimination model in assuming no prejudice or invidious motive by employers or employees, but rather that employers use average characteristics of groups to predict individual worker attributes.[2] The early models examined only the distributive consequences of statistical discrimination (see Kenneth Arrow, 1972, 1973; Edmund Phelps, 1972; Dennis Aigner and Glen Cain, 1977; and George Borjas and Matthew Goldberg, 1978). The literature tacitly assumed that statistical discrimination, whereby firms use valid (and free) information solely to maximize profits, must be more efficient than an economy where firms ignore the information. The standard statistical-discrimination model thus presents society with an uncomfortable tradeoff. In prohibiting statistical discrimination, society must accept lower national output.

Recently, Shelly Lundberg and Richard Startz (1983) expanded the standard model of Aigner-Cain to examine some of the efficiency effects of statistical discrimination. Aigner and Cain had considered a labor market where employers can only imperfectly test worker ability, and the test predicts more accurately for workers of one group than for another. Aigner and Cain had derived from this model an employer wage scale that offers a worker a convex combination of the mean productivity of his group and his individual test score, and had examined the distributional consequences between groups of such an offer. Lundberg and Startz examined how these wage offers might affect human capital decisions. They first showed that workers in such a world invest too little in training, because individual workers are not completely compensated for individual increases in productivity. Lundberg and Startz then concluded that statistical discrimination may exacerbate the inefficiencies. Although the wage of the favored group rises closer to the allocatively efficient wage, the wage of the disfavored group falls further from the efficient wage.

*Cornell Law School, Ithaca, NY 14853. This paper is a substantially revised section of my doctoral dissertation. I thank Peter Steiner and my dissertation committee members Frank Stafford, Glenn Loury, Theodore St. Antoine, and Gavin Wright for their guidance. I also acknowledge comments by Ron Ehrenberg, Olivia Mitchell, and other participants at the Labor Economics Workshop of Cornell's New York State School of Industrial and Labor Relations, and an anonymous referee.

[1]A third basic framework, somewhat outside the neoclassical tradition, uses labor market segmentation or "dual" labor markets to explain the occurrence and persistence of discrimination.

[2]I sometimes refer to statistical discrimination as a situation where an employer acts on a "true stereotype." Examples of stereotypes include "blacks are less skilled than whites," "women quit more frequently than men," and "women live longer than men." The term stereotype reflects society's moral distaste for statistical discrimination, under the battle cry "judge me, not my group." Yet the word "true" is a crucial modifier in the phrase true stereotype. The employer, I assume, responds only to correct group information (statements that are indeed true on average) because competitive forces will eliminate employer decisions based on false information. See my dissertation for an extended legal and moral evaluation of statistical discrimination.

In the present paper, I likewise examine the efficiency effects of statistical discrimination. I start from a different strain in the imperfect-information literature, the George Akerlof (1970) and Hayne Leland (1979) "lemons" model. Adapting this model to the labor market, I examine whether employers' use of group information will decrease allocative inefficiencies in labor supply. As Lundberg and Startz found for human capital investments, I find that statistical discrimination increases efficiency of labor supply for the favored group but decreases efficiency for the disfavored group. My general model outlines the parameters and suggests that the net efficiency effect cannot be determined a priori. Importantly, my model suggests that statistical discrimination *can* reduce the efficiency of the economy even if the two groups differ in their underlying productivities. In Section III, a variant of the model more strongly concludes that statistical discrimination *will* exacerbate inefficiencies under certain conditions.

I. Labor Supply Distortions from Limited Information

Statistical discrimination would not occur if employers knew the ability of individual workers. In many settings, however, employers cannot obtain individual information, so they resort to less precise group information to fill the information void. To determine whether "filling the void" increases allocative efficiency, I first examine the efficiency losses from imperfect information. In Section II, I give employers group information to fill the void.

A. *Labor Demand*

Employers cannot distinguish among individuals, so they set the wage rate at the average product of employed workers. For simplicity, I assume that output equals the amount of effective labor used, uF, where u is the average ability of employed workers and F is the number of employed workers. In a competitive market, then, the demand relation for labor quality is

$$u = w. \tag{1}$$

B. *Labor Supply*

Individuals can work in one of two markets. In the "standardized" job market described above, employers cannot distinguish among workers. All workers receive the same wage. The alternate market—the "individualized" market—can identify and pay workers according to individual productivity.[3] This market includes self-employed persons (including homemakers), persons paid by piece work, entrepreneurs, and managers for whom the bottom line reflects individual managerial ability. The key distinction is that employers in the standardized market have an incentive to discriminate statistically, whereas employers in the individualized market know individual ability and thus need not rely on less precise group information.[4]

A critical assumption is that the more skilled workers in the standardized market can also produce more in the individualized market.[5] As the standardized wage increases, more-able workers are drawn into this market. The last worker necessarily has a higher marginal product than the average worker (and workers in the individualized market still higher marginal products).

To model this more formally, let a be an index over the interval $(0,Z)$ of an individual's productivity in the standardized market. Let $f(a)$ be the number of workers of ability a. A worker supplies one unit if he decides to work in the standardized market. A person will work if the standardized-market wage, w, exceeds his individualized-market productivity, P. To capture the idea that more-able standardized workers can also produce more in the individualized market, I assume the

[3] To avoid general equilibrium problems of shifting prices, I assume that workers produce the same product in the standardized and individualized market and that wages are paid in output product units.

[4] See Michael Spence (1974a, b) for a discussion of distinctions between markets where employers can observe individual productivity and markets where signalling and statistical discrimination will occur.

[5] The assumption is equivalent to Akerlof's and Leland's assumption that as the market price of goods rises, higher quality goods are sold. See also James Heckman (1974), who deals with the wage/quality issue in the labor market.

following relation:

$$(2) \qquad P = P(a) \quad P'(a) > 0, \quad P''(a) > 0.$$

The positive second derivative suggests that, at some point, individual productivity in a collective setting has some limits.

Define A as the highest ability level appearing in the standardized market. Given the labor supply assumption, we have $A = A(w)$, where $A(w)$ is the inverse function of P. The number of workers willing to work in the standardized market for a given wage is

$$(3) \qquad F(w) = \int_0^{A(w)} f(a)\, da.$$

Average ability in the standardized market, u, increases with w:

$$(4) \qquad u(w) = \int_0^{A(w)} af(a)\, da / F(w).$$

Note that the supply equations for the quantity and quality of labor (equations (3) and (4)) are dependent on each other.

C. *Suboptimal Equilibrium*

Substituting equations (3) and (4) into equation (1) creates a single equilibrium equation in w:[6]

$$(5) \qquad u(w) = w.$$

Figure 1 graphs the equilibrium wage and

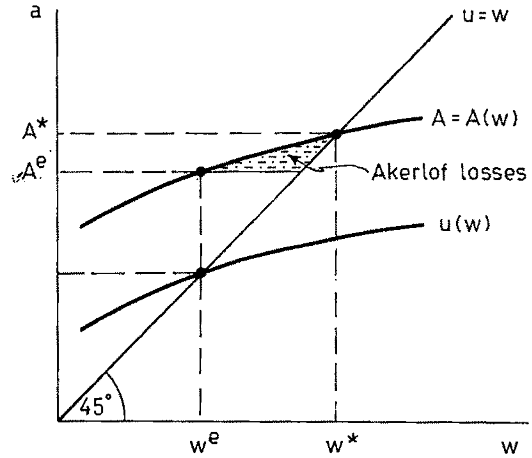


FIGURE 1. EFFICIENCY LOSSES WITH NO INFORMATION ABOUT INDIVIDUAL WORKER ABILITY

average ability, w^e and u^e. Overall social product is the sum of production in the standardized and individualized market (call this *GNP*):

$$(6) \quad GNP = \int_0^{A(w)} af(a)\, da + \int_{A(w)}^{Z} P(a) f(a)\, da.$$

GNP would be maximized by allocating workers to equate the marginal output of standardized and individualized market workers (point A^*, w^* in Figure 1), with all workers of ability less than A^* working in the standardized market. As can be seen from Figure 1, an unregulated economy allocates too few workers to the standardized market.[7] The problem is that firms value the (above average) marginal worker as a person of average ability, because firms must treat all persons on average. This discourages able

[6] Stability requires that $u(w)$ be concave in the relevant range, a condition I assume throughout the paper. Differentiating equation (4), and dropping the arguments of the functions for brevity, yields $u'(w) = fA'(A-u)/F > 0$. Differentiating again yields $u''(w) = fA''(A-u)/F + [A']^2 f'(A-u)/F + fA'(A'-u)/F + [fA']^2(u-A)/F^2$. The last term is always negative. Term 1 is negative since, by the second-order restrictions of equation (2), $A'' < 0$. Term 2 is negative if $f' < 0$, which occurs for all unimodal distributions whenever A exceeds the mode. Term 3 is negative if $f(A-u) > F$, which may well not be true. Although there is some indeterminateness for the sign of u'' in general, u is concave for a wide range of ability distributions. This holds, for example, for all ability distributions of the form $f(a) = ca^n$, $n > -1$ or < -2, since $u'' = A''(n+1)/(n+2)$.

[7] The government could achieve first-best social efficiency by subsidizing the wages of all workers, with an optimal subsidy being $S^* = A^* - u^*$. This possibility is analogous to the government policy mentioned by Lundberg and Startz (p. 344, fn. 6) of subsidizing wages to induce optimal human capital investment by workers. As they point out, in practice it would be extremely difficult to calculate and implement this subsidy.

persons from working in the standardized market. Indeed, in equilibrium, the marginal worker can produce in the individualized market what the average person produces in the standardized market. The problem of limited information is thus the gap between the marginal contributions of the average and marginal worker.[8] Following the original lemons model, I call the losses caused by this gap "Akerlof losses." (See Figure 1.)

II. Adding Statistical Discrimination to the Model

Suppose now that firms in the standardized market know something about individual workers. Our question is whether this information will reduce the Akerlof inefficiencies. To create a statistical-discrimination model, I give firms additional information in the form of a true stereotype.

Workers belong to one of two groups, which firms can costlessly identify. Within each group, workers vary in ability. All workers of ability a have the same labor market behavior, regardless of the group to which they belong. For reasons exogenous to the model, however, workers of group 1 have a higher average ability in the standardized market than workers of group 2, for any given wage. The true stereotype, then, is that "group 1 workers are more productive than group 2 workers."

As before, production in the standardized market equals the amount of effective labor, $L = u^1F^1 + u^2F^2$. Firms using the group information will treat all persons within a group equally, and will offer a wage of $w^1 = u^1$ to group 1 persons, and a lower wage of $w^2 = u^2$ to group 2 persons. The quantity and quality of labor supplied for each group are analogous to equations (3) and (4). Figure 2 graphs the separating equilibrium.

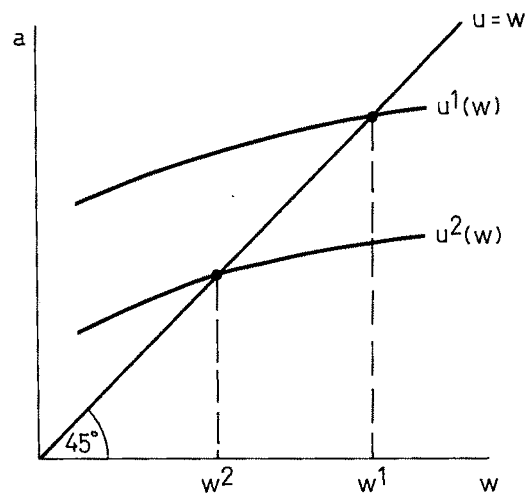


FIGURE 2. EQUILIBRIUM WAGES WITH GROUP INFORMATION

Our question is whether statistical discrimination increases output. If firms use group information, GNP^d (GNP discriminatory) is

$$(7)\quad GNP^d = \int_0^{A^1 = A(w^1)} af^1(a)\,da + \int_0^{A^2 = A(w^2)} af^2(a)\,da$$

$$+ \int_{A^1}^{Z} P(a) f^1(a)\,da + \int_{A^2}^{Z} P(a) f^2(a)\,da.$$

Subtracting (7) from (6), statistical discrimination increases social output if and only if

$$(8)\quad \int_A^{A^1} [a - P(a)]\, f^1(a)\,da > \int_{A^2}^{A} [a - P(a)]\, f^2(a)\,da.$$

As equation (8) shows, the efficiency of statistical discrimination depends on three factors. First is the relative number of group 1 and group 2 persons who are highly skilled, marginal workers. Second is the extent of the shift in wages. Third is the net change in social output as group 1 workers enter the standardized market and group 2 workers leave for the individualized market. The issue is whether the gap in the marginal worker's production in the two markets rises or falls with ability over the relevant range. In terms

[8] In a similar vein, Spence (1975) has analyzed the problem of the gap between the valuation of quality by the average and marginal consumer.

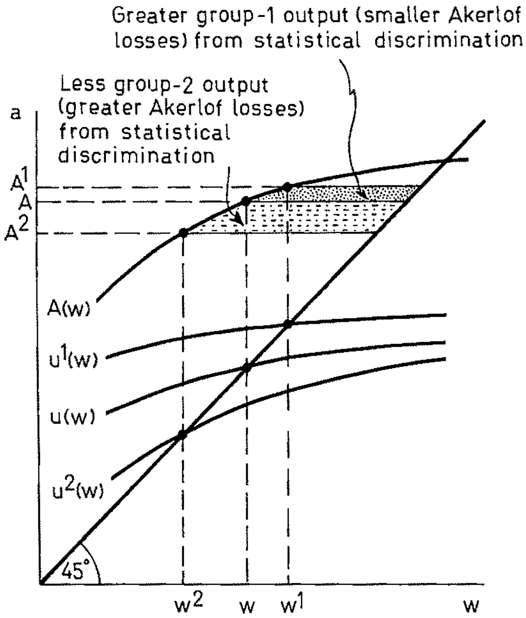


FIGURE 3. EFFICIENCY COMPARISON OF DISCRIMINATORY AND NONDISCRIMINATORY EQUILIBRIA

of the model, this third factor will tend to make statistical discrimination efficient if $P' > 1$.

Figure 3 sketches the gains and losses in efficiency from the second and third factors. The $A(w)$ schedule, showing the ability of the most-able standardized-market worker for a given wage, is identical for both groups since workers of the same ability behave identically regardless of group. It is along this $A(w)$ curve of marginal workers that the relative efficiency calculations are made. But Figure 3 also shows that the wages of the two groups, which determine the extent of the relative Akerlof losses, are based on non-marginal group averages. The wages thus have no necessary relation to efficiency. A ban on statistical discrimination pushes the wage of group 2 workers closer to a socially optimal incentive to work, but causes the wage of group 1 workers to diverge further from an efficient level. The net efficiency effect depends on the shape of the ability functions of the two groups. As drawn, Figure 3 suggests that statistical discrimination decreases output. The large loss in group 2 average ability outweighs the increase in group 1 average ability. Note that the only restriction on the underlying ability distributions is that u^1 exceed u^2 over the relevant range. Thus, for example, statistical discrimination can be inefficient even if the groups differ in the underlying average ability.

Statistical discrimination is most likely to be inefficient when the disfavored group has relatively large numbers of unskilled workers, holding down the average ability of the group, while the skilled workers are more evenly dispersed between groups. Figure 3 was drawn with this distribution in mind: the average abilities of the two groups at low levels of the truncated distribution (u^1 and u^2 at low wages) vary widely, but the gap in average ability narrows as more-skilled workers enter the labor market.

III. The Inefficiency of Statistical Discrimination when Labor Supply Elasticity Differs by Group

The previous model treated the standardized/individualized market choice of workers as invariant across groups. The group differences that employers observed arose from differences in ability distributions ($f^1 \neq f^2$). Group differences can also occur when the opposite assumptions are made: standardized ability distributions are the same for groups 1 and 2 but the reservation-wage functions (indicating individualized-market productivity) differ by group.

Suppose the standardized labor supply of group 1 is highly inelastic, so that virtually all members of group 1 work in the standardized market for any wage in the relevant range. In other words, A^1 and u^1 are nearly constant functions of the wage. Suppose the supply elasticity of group 2 is greater: indeed, assume that the supply response of group 2 persons is analogous to equations (3) and (4) of the previous model. In this situation, employers in the standardized market will observe that the average productivity of group 1 exceeds that of group 2, because virtually all group 1 persons appear in the standardized market but the most-able group 2 persons remain in the individualized market.

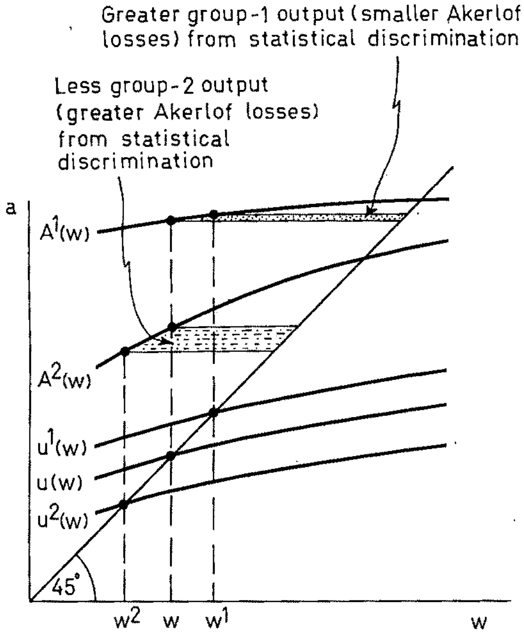


FIGURE 4. INEFFICIENCY WHEN LABOR SUPPLY ELASTICITY DIFFERS BY GROUP

Figure 4 graphs this variant of the model. The important result is that statistical discrimination is now *unambiguously inefficient.* Most group 1 workers will work regardless of the wage, so increasing their wage from w to w^1 does not induce a larger supply of workers into the standardized market. Lowering the group 2 wage from w to w^2, however, discourages talented persons from working in the standardized market, and thus lowers output. Society is more productive in this situation if firms ignore group information.

IV. Concluding Remarks

I have answered the question of the title of this piece with a definite "maybe not." Even though statistical discrimination is based on free, accurate information, it does not necessarily allocate resources more efficiently than if firms ignored the information. I have identified two situations in which firms' use of group information may exacerbate the labor supply distortions of limited information. First, one group may have many unskilled persons who lower the expected ability of all persons in the group. The skilled workers at the margin, however, may be more evenly dispersed between groups. If so, using group information to set wages will discourage more able workers than it encourages. Societal output decreases. Second, labor supply inefficiencies arise if the reservation-wage function differs between groups. In such a case, wages set by group information will discourage group 2 persons without an offsetting encouragement of group 1 persons.

To be certain of the inefficiency of statistical discrimination in specific situations, one needs empirical documentation. Unfortunately, empirical testing of statistical discrimination (and, indeed, of the signalling hypothesis in general) is difficult.[9] Further refinement of the models should enable empirical testing. In the meantime, the negative theorem remains important in shifting the burden of proof on the acceptability of statistical discrimination. Firms commonly assert that they must statistically discriminate to maximize profits, which—as Adam Smith showed—benefits society in general. The negative theorem indicates, however, that an a priori efficiency claim cannot be used to justify statistical discrimination. Statistical discrimination may indeed be socially inefficient. Perhaps we should thus be more confident in acting upon our moral approbation of it, at least until empirical documentation demonstrates in which direction the efficiencies lie.

[9] For one effort, see Daniel Dick and Marshall Medoff (1976).

REFERENCES

Aigner, Dennis and Cain, Glen, "Statistical Theories of Discrimination in Labor Markets," *Industrial and Labor Relations Review*, January 1977, *30*, 175–87.

Akerlof, George, "The Market for Lemons: Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, *84*, 488–500.

Arrow, Kenneth, "Some Mathematical Models of Race Discrimination in the Labor

Market," in Anthony Pascal, ed., *Racial Discrimination in Economic Life*, Lexington: Lexington Books, 1972.

______, "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton: Princeton University Press, 1973.

Becker, Gary, *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

Borjas, George and Goldberg, Matthew, "Biased Screening and Discrimination in the Labor Market," *American Economic Review*, December 1978, *68*, 918–22.

Dick, Daniel and Medoff, Marshall, "Filtering by Race and Education in the U.S. Manufacturing Sector: Constant-Ratio Elasticity of Substitution Evidence," *Review of Economics and Statistics*, May 1976, *58*, 148–55.

Heckman, James, "Shadow Wages, Market Wages, and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

Leland, Hayne, "Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards," *Journal of Political Economy*, December 1979, *87*, 1328–46.

Lundberg, Shelly and Startz, Richard, "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review*, June 1983, *73*, 340–47.

Phelps, Edmund, "The Statistical Theory of Racism and Sexism," *American Economic Review*, September 1972, *62*, 659–61.

Schwab, Stewart, "Stereotypes, Imperfect-Information Theories, and Statistical Discrimination in Labor Markets," unpublished doctoral dissertation, University of Michigan, 1981.

Spence, A. Michael, (1974a) *Market Signalling: Information Transfer in Hiring and Related Screening Processes*, Cambridge: Harvard University Press, 1974.

______, (1974b) "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *Journal of Economic Theory*, March 1974, *7*, 296–332.

______, "Monopoly, Quality, and Regulation," *Bell Journal of Economics*, Autumn 1975, *6*, 417–29.

Wage Indexation and the Effect of Inflation Uncertainty on Employment: An Empirical Analysis

By A. STEVEN HOLLAND*

In his Nobel Lecture, Milton Friedman (1977) argued that the greater uncertainty associated with higher inflation leads to a misallocation of resources because of shorter duration of contracts and reduced efficiency of the price system. The result is reduced economic growth and, possibly, more unemployment (i.e., a positively sloped Phillips curve) over the fairly long term. In a subsequent article, Maurice Levi and John Makin (1980) found a significant negative impact of inflation uncertainty on employment growth. Evidence of a similar nature was reported by Yakov Amihud (1981), Makin (1982), and Ronald Ratti (1985), while Donald Mullineaux (1980) found a significant positive effect of inflation uncertainty on the rate of unemployment and a negative effect on industrial production. Given the substantial body of empirical literature linking higher inflation to greater inflation uncertainty, this provides support for Friedman's hypothesis.[1]

Friedman also noted, however, that in the very long run, institutions should adapt to an inflationary economy in a way that offsets much of the real effect of higher inflation. An example of such adaptation is more widespread indexation of wages. Levi and Makin recognized the potential impact of indexing but did not attempt to estimate it: "To the extent that inflation uncertainty persists and causes lower employment, our results tend to support the case for a wider use of indexing of nominal contracts, which should reduce the impact of uncertainty felt on the real sector" (p. 1026).

The purpose of this article is to estimate the impact of inflation uncertainty on employment, while also considering the second-round effects of labor market adjustments designed to reduce the risk associated with inflation uncertainty. Despite the limited scope of the data, an increase in the prevalence of wage indexation in major collective bargaining contracts is taken to indicate a general increase in the responsiveness of nominal wages to inflation surprises.[2] In other words, as the percentage of contracts with indexation clauses increases, the degree to which already indexed wages adjust to price level changes is assumed to increase. Furthermore, the effect is assumed to extend beyond the sector of the labor market covered by major collective bargaining agreements to smaller union contracts and even to nonunionized labor.

This article proceeds as follows. Section I discusses the measurement of inflation uncertainty and the level of wage indexation and estimates the impact of inflation uncertainty on indexation in the United States for the period 1961–83. Section II examines the impact of inflation uncertainty, indexation, and unanticipated inflation on employment. Section III presents the results of simulations designed to illustrate how increased wage indexation offsets at least part of the adverse

*Department of Economics, University of Kentucky, Lexington, KY 40506. Helpful comments from R. W. Hafer, Ronald Ratti, Richard Sheehan, Daniel Thornton, two referees, and the participants in seminars at the Board of Governors of the Federal Reserve System, Claremont College, Georgia State University, and the University of Kentucky are gratefully acknowledged. This research was conducted at the Federal Reserve Bank of St. Louis with assistance from Jude Naes. The views expressed do not necessarily reflect those of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

[1] My 1984 article provides a review of the literature linking higher inflation to greater inflation uncertainty.

[2] Formal indexing typically applies only to contracts in the unionized sector—less than 25 percent of the U.S. labor market. This measure should serve the purpose at hand, however, since the behavior of union wages influences the wages of other workers, and since adjustments to greater inflation uncertainty in the unionized sector can be expected to occur at roughly the same time as adjustments in other sectors of the labor market.

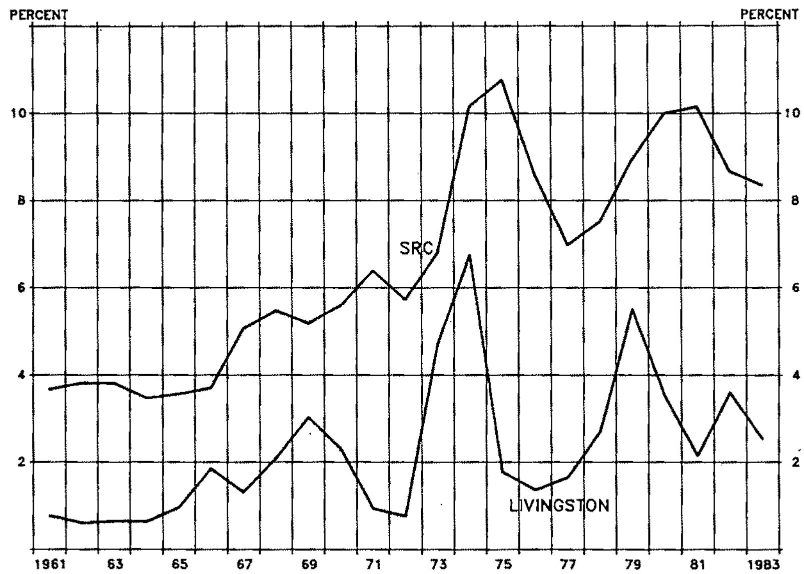


FIGURE 1

impact of inflation uncertainty on the level of employment in the long term.

I. The Effect of Inflation Uncertainty on the Level of Indexation

Various proxies for inflation uncertainty have been used in empirical research, including estimates of the variability of inflation rates, the cross-sectional variability of expected inflation rates and the variability of inflation forecast errors. The latter is most closely related to the concept of uncertainty. Alex Cukierman and Paul Wachtel (1982) suggest the mean squared error (MSE) of inflation forecasts from a survey as the best available measure. This is a cross-sectional measure of the dispersion of forecast errors across individuals:

$$(1)\quad MSE_t = \frac{1}{n}\sum_{i=1}^{n}(\pi_{it}^e - \pi_t)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\pi_{it}^e - \bar{\pi}_t^e)^2 + (\bar{\pi}_t^e - \pi_t)^2,$$

where n is the number of forecasters, π is the actual rate of inflation, π_i^e is the rate of inflation forecast by individual i, and $\bar{\pi}^e$ is the mean forecasted rate of inflation among respondents. The first term on the right-hand side of the equation is the variance of individual forecasted values, and the second term is the squared mean forecast error. The square root of the first term (the standard deviation of the forecasts) is the measure most often used in previous studies.[3]

Figure 1 plots the root mean squared error ($RMSE$) of year-end to year-end inflation forecasts from both the Livingston and the University of Michigan SRC surveys.[4] The

[3]The empirical tests presented herein were also performed using the standard deviation of inflation forecasts. Where there are significant discrepancies between this set of results and those reported in the text, it will be noted.

[4]The survey by Joseph Livingston of *The Philadelphia Inquirer* asks respondents to predict a number of economic indicators including the Consumer Price Index (CPI). The year-end to year-end forecasts used here are actually 14-month forecasts since respondents are thought to know only the level of the October CPI when they turn in their prediction in December of the level of the CPI for the following December. With this in mind, John Carlson (1977) has revised Livingston's data on inflation expectations, and this revised data (updated through 1983) is used here. The Survey Re-

two series are highly correlated—the correlation coefficient is 0.61—but their magnitudes differ substantially. The *RMSE* of the SRC survey, which is a broad survey of households, is much higher than that of the Livingston survey, which is a narrower survey of professional economic forecasters. Both series are substantially higher in the 1970's and 1980's than in the 1960's. The average level of *RMSE* for the Livingston survey for 1961–72 is 1.32 and for 1973–83 is 3.29. For the SRC survey, these averages are 4.62 and 8.81, respectively.[5]

The proxy for the responsiveness of nominal wages to inflation surprises is the number of workers covered by cost-of-living adjustment (COLA) clauses as a percentage of the total number of workers subject to major collective bargaining contracts.[6] This variable is referred to below as the level of wage indexation.

An increase in inflation uncertainty is expected to cause an increase in the level of indexation. Because his real wage is more uncertain, a risk-averse worker whose nominal wage is fixed by contract for a specified period suffers a reduction in utility from greater inflation uncertainty. Indexing offsets this decline in utility, however, because it helps to insulate the worker's real wage from the effects of unanticipated inflation. A similar analysis applies to firms that are risk averse: indexing reduces the impact of lower-than-anticipated inflation on the real cost of labor inputs.[7] Other factors that could affect the level of wage indexation such as the cost of providing indexation, the degree to which nonlabor income is indexed, and the power of labor unions are assumed to be constant during the period of analysis. It should be noted, however, that if there is a significant cost to indexing, then individuals must interpret at least part of an increase in inflation uncertainty as persistent if they are to adjust the level of indexation.

The level of wage indexation is assumed to be determined by the following process:

$$(2)\qquad I_t - I_{t-1} = \lambda(I_t^* - I_{t-1}) + \varepsilon_{1t},$$

$$(3)\qquad I_t^* = \alpha_0 + \sum_{j=0}^{m} \alpha_j U_{t-j},$$

where I is the level of indexation, I^* is the desired level of indexation, U is inflation uncertainty, and ε_1 is a random error term. Partial adjustment of the current level of indexation to its desired level results from the presence of multiyear contracts. Combining (2) and (3), the equation to be estimated is

$$(4)\qquad I_t = \alpha_0\lambda + \lambda \sum_{j=0}^{m} \alpha_j U_{t-j} + (1-\lambda)I_{t-1} + \varepsilon_{1t}.$$

Table 1 presents the results of ordinary least squares regressions for the period 1961–83 using the annual data for the percentage of workers with COLA clauses and both of the inflation uncertainty measures presented in Figure 1.[8] Four lagged values of

search Center (SRC) survey asks households to predict the rate of change in their cost of living over the next 12 months. I use the fourth-quarter to fourth-quarter forecasts. Prior to 1966, respondents to the SRC survey were not asked to predict a specific rate of inflation, but were asked to give a general impression of the future behavior of prices. F. Thomas Juster and Robert Comment (1979) have transformed the pre-1966 data to provide a series consistent with the post-1966 data.

[5]Note that the highest values of *RMSE* for each survey occur for the years during or immediately following large energy shocks. In my earlier paper, I found that both higher inflation and energy shocks affected the *RMSE* of the Livingston forecasts. It should be noted that the generally higher values of *RMSE* for the SRC survey are due primarily to a higher variability of responses and not to larger forecast errors.

[6]The data on cost-of-living adjustments from 1960–70 come from Wallace Hendricks and Lawrence Kahn (1983). For 1971–83, the data are from the *Monthly Labor Review*. Major collective bargaining agreements are those that apply to 1,000 or more workers.

[7]For a more detailed analysis of the impact of inflation uncertainty on wage indexation, see Leif Danziger (1984). JoAnna Gray (1976) presents a model that implies that in order to minimize the deviation of output from full-information output, wage indexation responds positively to monetary variability, *ceteris paribus*.

[8]Equation (4) should properly be considered a linear approximation of a nonlinear model, because I cannot be greater than 100 (percent). Therefore, the estimates

TABLE 1—THE IMPACT OF INFLATION UNCERTAINTY ON WAGE INDEXATION, 1961–83[a]

	Livingston Survey (1)	SRC Survey (2)
Constant	1.69	−3.03
	(0.62)	(−1.60)
U_{t-1}	1.30	1.62
	(1.71)	(2.00)
U_{t-2}	2.13	1.79
	(2.31)	(1.28)
U_{t-3}	0.82	−1.00
	(0.89)	(−0.72)
U_{t-4}	1.30	2.12
	(1.40)	(2.22)
Sum	5.55	4.53
	(3.56)	(6.95)
I_{t-1}	0.67	0.39
	(6.36)	(4.19)
$\bar{R}^2$	0.92	0.96
$S.E.$	4.50	2.93
Dh	0.33	0.76

[a]The t-statistics are shown in parentheses; Dh refers to Durbin's h-statistic for testing for autocorrelation in regressions with lagged dependent variables.

the $RMSE$ of the inflation forecasts, the proxy for U, have a significant positive impact on wage indexation in each regression.[9] The contemporaneous term was dropped from each equation because its impact was negligible.[10] One year's lag of the dependent variable was significant in each regression, which implies that the effect of inflation uncertainty on indexation lasts beyond four years due to partial adjustment. This lagged effect is consistent with Friedman's notion that it takes time for institutions to respond to greater inflation uncertainty.[11]

At first glance, the magnitude of the long-term effect of inflation uncertainty on the level of indexation differs substantially between the two estimates. For the Livingston data, the cumulative effect of a permanent increase in U of one percentage point is an increase in I of 17.1 percentage points, while for the SRC survey, a permanent increase in U of one percentage point leads to an increase in I of only 7.4 percentage points. Recall, however, that the scales of the two variables differ substantially, and that a two-percentage-point increase in the average level of the Livingston measure occurred simultaneously with a four-percentage-point increase in the average level of the SRC measure. Therefore, their effects on the level of indexation are actually quite similar.

II. The Effects of Inflation Uncertainty and Wage Indexation on Employment

A model that considers the relationships between inflation uncertainty, indexation of labor contracts and employment has been developed by Amihud. He shows that in the absence of wage indexation, risk-averse laborers who maximize the utility from leisure and real consumption reduce the supply of labor services offered to the market in the face of greater inflation uncertainty, as long as the substitution effect between leisure and consumption dominates the income effect—that is, the labor supply function slopes upward. This reduction in labor supply also occurs if an individual's nonlabor income can be hedged against unexpected price level movements, while his labor income cannot. With either risk-neutral or risk-averse business firms, the overall effect is reduced employment, since the demand for labor either

apply to levels of U in the neighborhood of those experienced during the sample period.

[9]Selection of the number of lags used in these and other regressions in the paper are based on standard t- and F-tests. In cases in which the number of lags chosen by this method would differ between the two sets of data, however, the longer of the two lag lengths is chosen so that the estimates might be more directly comparable. In no case is there a large discrepancy between the results with a shorter lag length and those presented in the text. Furthermore, in each regression, the F-test rejects the hypothesis that all of the coefficients of the lagged values are zero at the 5 percent level of significance.

[10]The value of the t-statistic for the contemporaneous value of U was less than 0.5 in each regression.

[11]If the standard deviation of the forecasts is used as the measure of inflation uncertainty instead of the $RMSE$, the effects are similar. The differences are that the length of the lagged effect is shorter, and autocorrelation is present in the regression using the Livingston data.

remains the same or falls. On the other hand, if there is 100 percent indexing, inflation uncertainty has no effect on the supply and demand for labor.

In this section, I test the hypothesis that a higher level of wage indexation reduces the size of the adverse impact of inflation uncertainty on employment. Changes in inflation uncertainty (U) and indexation (I) affect the equilibrium level of employment, as specified by

$$(5)\quad \log E_t^* = \beta_0 + \beta_1 t - \sum_{k=0}^{p} \delta_k U_{t-k} + \sum_{r=0}^{s} \phi_r I_{t-r},$$

where $\log E^*$ is the natural logarithm of the equilibrium level of employment and t is a time trend. The actual level of employment (E) is determined according to the following:

$$(6)\quad \log E_t = \log E_t^* + \theta(1 - I_t/100)(\pi_t - \pi_t^e) + \varepsilon_{2t},$$

where $\pi - \pi^e$ is the unanticipated rate of inflation and ε_2 is a random error term. Unanticipated inflation causes the actual real wage rate to deviate from its equilibrium level in the short term, which, depending on the responses of the demand and supply for labor, may cause employment to also deviate from its equilibrium level. The direction of this effect is not specified a priori, since it depends on whether the forces of labor demand or labor supply predominate in disequilibrium situations.[12] A reduction in the real wage causes the quantity of labor demanded to increase, but may also cause the quantity of labor supplied to decrease. This paper postulates, however, that any effect of an inflation surprise on employment is smaller the higher is the level of indexation (I), because indexing mitigates the effect of inflationary shocks on real wages. Thus, a term accounting for the interaction between $\pi_t - \pi_t^e$ and I_t is included in equation (6).[13]

Combining (5) and (6), the equation to be estimated is

$$(7)\quad \log E_t = \beta_0 + \beta_1 t - \sum_{k=0}^{p} \delta_k U_{t-k} + \sum_{r=0}^{s} \phi_r I_{t-r} + \theta(1 - I_t/100)(\pi_t - \pi_t^e) + \varepsilon_{2t}.$$

Table 2 presents the results for the sample

TABLE 2—THE IMPACTS OF INFLATION UNCERTAINTY AND WAGE INDEXATION ON THE LOG OF EMPLOYMENT, 1961–83[a]

	Livingston Survey (1)	SRC Survey (2)
Constant	415	418
	(520)	(227)
t	2.16	2.23
	(25.86)	(13.77)
U_t	−0.66	−0.07
	(−2.66)	(−0.23)
U_{t-1}	−0.51	−1.02
	(−3.48)	(−2.45)
U_{t-2}	−0.25	−0.08
	(−1.77)	(−0.21)
U_{t-3}	−0.36	−0.42
	(−2.29)	(−1.37)
Sum	−1.78	−1.59
	(−4.50)	(−2.86)
I_t	0.10	0.17
	(3.12)	(3.24)
$(\pi_t - \pi_t^e)(1 - I_t/100)$	1.36	0.37
	(4.30)	(2.21)
ρ	0.39	0.70
	(2.01)	(4.64)
$\bar{R}^2$	0.997	0.996
$S.E.$	0.76	0.91
$D\text{-}W$	2.03	1.92

[a] The t-statistics are shown in parentheses; ρ is the estimated autocorrelation coefficient.

[12] See Cukierman (1980). If employment in disequilibrium is determined as the minimum of labor demand and labor supply—the "short-end" rule—then (6) should include positive and negative values of unanticipated inflation as separate regressors. When the equation was estimated in this form, none of the estimates of the effect of inflation uncertainty and indexation were affected, but the unanticipated price change variables were generally not statistically significant.

[13] The variable I is divided by 100 so that it is in ratio rather than percentage form. A potential complication is that indexed wage adjustments generally occur after an inflationary shock has occurred. The adjustment appears to take place within a year's period, however, since the lagged value of the variable $(1 - I/100)(\pi - \pi^e)$ was not statistically significant in any of the estimations.

period, 1961–83, using the Cochrane-Orcutt adjustment for autocorrelation.[14] For each set of data, the measures of inflation uncertainty have a significant negative impact on employment over the current and three previous years. The effect of the contemporaneous level of wage indexation is positive and statistically significant in both equations, though it is larger in the one using the SRC data.

Unanticipated inflation has a positive and significant impact on employment in each regression, an effect that has been restricted to grow smaller as the level of wage indexation rises. The size of this coefficient is substantially higher in the Livingston regression than in the SRC regression. Since unanticipated inflation reduces the real wage and increases the quantity of labor demanded, this indicates that labor demand effects are stronger than labor supply effects in disequilibrium.[15] Unfortunately, when the equations are run in unrestricted form, with separate terms for unanticipated inflation and the interaction between indexation and unanticipated inflation, multicollinearity is present.[16] When unanticipated inflation is included in each equation without any interaction with the level of indexation, its effect is smaller, but still positive and significant, and the other coefficients are not greatly affected.

The regressions in Table 2 were also reestimated in first difference form, providing equations for the growth rate of employment. The results, presented in Table 3, are nearly identical to those for the log of employment, except that the two estimates of the impact of indexation are closer to each other than they were in Table 2.[17]

TABLE 3—THE IMPACTS OF INFLATION UNCERTAINTY AND WAGE INDEXATION ON THE GROWTH RATE OF EMPLOYMENT, 1961–83[a]

	Livingston Survey (1)	SRC Survey (2)
Constant	2.07	2.25
	(15.28)	(11.38)
ΔU_t	−0.63	−0.18
	(−2.63)	(−0.68)
ΔU_{t-1}	−0.49	−0.92
	(−3.79)	(−2.36)
ΔU_{t-2}	−0.23	−0.16
	(−1.74)	(−0.48)
ΔU_{t-3}	−0.38	−0.37
	(−2.64)	(−1.28)
Sum	−1.73	−1.63
	(−4.93)	(−3.29)
ΔI_t	0.13	0.16
	(4.00)	(4.19)
$\Delta[(\pi_t - \pi_t^e)(1 - I_t/100)]$	1.36	0.37
	(4.10)	(2.31)
ρ	−0.38	−0.20
	(−1.98)	(−0.99)
$\bar{R}^2$	0.70	0.62
$S.E.$	0.83	0.93
$D\text{-}W$	1.65	1.82

[a]The t-statistics are shown in parentheses; ρ is the estimated autocorrelation coefficient.

Despite the similarity in the numerical estimates of the impacts of inflation uncertainty and wage indexation on both employment and employment growth between the two sets of data, the estimates have different implications. If each measure of uncertainty were to increase by the amount that occurred on average between the periods 1961–72 and 1973–83, the effect of both inflation uncertainty and indexation on employment would

[14]Total employment of the civilian labor force (seasonally adjusted) for the fourth quarter is the measure of E. The dependent variable is actually the log of E multiplied by 100.

[15]This result is also consistent with employment being completely demand determined in disequilibrium, as specified by Gray and by Stanley Fischer (1977). Furthermore, it is consistent with the implications of a model incorporating rational expectations and market clearing, such as that of Robert Lucas (1973), in which unanticipated inflation causes employment to rise above its "natural" level.

[16]The equations were also run in an even less restrictive form in which separate coefficients for the interaction between indexation and the actual inflation rate and between indexation and the expected inflation rate were estimated. This regression also displayed multicollinearity.

[17]The only difference between the results using the standard deviation of the forecasts and those reported in the text is a slightly shorter lag length for U. None of the results in this paper were altered when either dummy variables to account for the possible effects of energy shocks in 1973–75 and 1979–81, or an estimate of the relative price of energy were added to the regressions.

TABLE 4—SIMULATION RESULTS: LEVEL OF WAGE INDEXATION AND PERCENTAGE DEVIATION OF EMPLOYMENT FROM TREND

	Livingston Survey		SRC Survey	
Year	Indexation (I)	Percentage Deviation of Employment (E)	Indexation (I)	Percentage Deviation of Employment (E)
0	20.0	0.00	20.0	0.00
1	20.0	−1.31	20.0	−0.27
2	22.6	−2.06	26.5	−3.18
3	28.6	−1.98	36.2	−1.85
4	34.3	−2.15	35.9	−3.54
5	40.8	−1.53	44.4	−2.13
6	45.1	−1.11	47.6	−1.57
7	48.0	−0.83	48.9	−1.36
8	50.0	−0.64	49.4	−1.27
9	51.3	−0.51	49.6	−1.24
10	52.2	−0.43	49.7	−1.22
11	52.8	−0.37	49.7	−1.22
12	53.2	−0.33	49.7	−1.22
13	53.5	−0.30	49.7	−1.22
14	53.7	−0.28	49.7	−1.22
15	53.8	−0.27	49.7	−1.22
16	53.9	−0.26	49.7	−1.22
17	54.0	−0.26	49.7	−1.22
18	54.0	−0.25	49.7	−1.22
19	54.0	−0.25	49.7	−1.22
20	54.1	−0.25	49.7	−1.22

be much greater in the SRC regressions than in the Livingston regressions. This will be illustrated more clearly in the next section.

III. Simulation Results: The Long-Term Impact of Inflation Uncertainty on Employment

The evidence presented above suggests that a permanent increase in inflation uncertainty causes the level of employment to fall below its constant growth rate trend, but a subsequent rise in wage indexation causes it to move back toward trend. In this section, simulations are performed using the estimates from Tables 1 and 2 to illustrate this point. Although the simulation period is 20 years, the simulated values beyond the first few years are not meant to be taken as precise estimates.

In the first simulation, the root mean squared error of the Livingston inflation forecasts increases permanently in year 1 by 2.0 percentage points from a steady state in which the growth rate of employment is 2.16 percent (the coefficient for the time trend in col. 1, Table 2) and the percentage of contracts with indexation (I) is chosen to be 20 percent. This simulated increase in *RMSE* is approximately the difference in the mean value of the *RMSE* from the Livingston survey between 1961–72 and 1973–83, and the starting value of I is approximately the level of the mid-1960's. The simulated values of I and of the percentage deviation of the level of employment (E) from its constant growth rate trend are presented in the first two columns of Table 4. The percentage of contracts with indexation rises to almost 29 percent in year 3 and continues to increase at a declining rate until it is just over 54 percent. The level of employment is about 2 percent below trend in years 2, 3, and 4, and then increases toward trend in subsequent years. In the long term, the permanent rise in inflation uncertainty results in E being about 0.25 percent below what it would otherwise have been. The adjustments are essentially complete by year 18. Thus, Friedman's no-

tion that it could take a number of years for the economy to adjust to greater inflation uncertainty is borne out by these simulation results.

Simulations using the *RMSE* from the SRC survey lead to a similar pattern of responses to an increase in inflation uncertainty, but different magnitudes of the effects. The simulated effects of a permanent increase in *RMSE* of 4.0 percentage points are presented in columns 3 and 4 of Table 4. The use of a simulated increase of 4.0 instead of 2.0 percentage points in *RMSE* is due to the different magnitudes of the two measures of inflation uncertainty. As before, the simulation approximates recent changes in the average level of *RMSE*. The percentage of contracts with indexation increases more rapidly initially than in the other simulation and takes less time to approach its new equilibrium level; it arrives at the new equilibrium value of 49.7 percent in year 10. The deviation of employment from trend is greater in this simulation than in the other one. It reaches a maximum of 3.54 percent in year 4 and then falls to 1.22 percent by year 10. Therefore, the specification using the SRC data shows both larger initial and long-term effects of inflation uncertainty on employment and a shorter adjustment period than the specification using the Livingston data. The basic result is the same, however: the impact of wage indexation offsets a large part, but not all, of the effect of a permanent change in inflation uncertainty on the level of employment.

IV. Conclusions

The evidence presented here indicates that a permanent increase in inflation uncertainty has a depressing influence on employment initially, but then the labor market adapts in a way that offsets most of the effect in later periods. The mechanism through which this occurs is a greater degree of adjustment of nominal wages to inflationary shocks (proxied by a higher percentage of indexed labor contracts in major collective bargaining agreements), which increases the labor services provided by risk-averse laborers and the demand for labor by risk-averse business firms. To the extent that higher inflation is associated with greater inflation uncertainty, these findings support Friedman's notion that a positively sloped Phillips curve exists over a fairly long interim period, but in the very long run, the economy moves back toward a natural rate of unemployment.

REFERENCES

Amihud, Yakov, "Price-Level Uncertainty, Indexation and Employment," *Southern Economic Journal*, January 1981, *47*, 776–87.

Carlson, John A., "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, Winter 1977, *6*, 27–56.

Cukierman, Alex, "The Effects of Wage Indexation on Macroeconomic Fluctuations: A Generalization," *Journal of Monetary Economics*, April 1980, *6*, 147–70.

______ **and Wachtel, Paul,** "Inflationary Expectations: Reply and Further Thoughts on Inflation Uncertainty," *American Economic Review*, June 1982, *72*, 508–12.

Danziger, Leif, "Stochastic Inflation and Wage Indexation," *Scandinavian Journal of Economics*, No. 3, 1984, *86*, 326–36.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Friedman, Milton, "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, June 1977, *85*, 451–72.

Gray, JoAnna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

Hendricks, Wallace E. and Kahn, Lawrence M., "Cost-of-Living Clauses in Union Contracts: Determinants and Effects," *Industrial and Labor Relations Review*, April 1983, *36*, 447–60.

Holland, A. Steven, "Does Higher Inflation Lead to More Uncertain Inflation?," *Federal Reserve Bank of St. Louis Review*, February 1984, *66*, 15–26.

Juster, F. Thomas and Comment, Robert, "A Note on the Measurement of Price Expectations," Working Paper, Institute for Social Research, University of Michigan, 1979.

Levi, Maurice D. and Makin, John H., "Inflation

Uncertainty and the Phillips Curve: Some Empirical Evidence," *American Economic Review*, December 1980, *70*, 1022–27.

Lucas, Robert E., Jr., "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

Makin, John H., "Anticipated Money, Inflation Uncertainty and Real Economic Activity," *Review of Economics and Statistics*, February 1982, *64*, 126–34.

Mullineaux, Donald J., "Unemployment, Industrial Production, and Inflation Uncertainty in the United States," *Review of Economics and Statistics*, May 1980, *62*, 163–69.

Ratti, Ronald A., "The Effects of Inflation Surprises and Uncertainty on Real Wages," *Review of Economics and Statistics*, May 1985, *67*, 309–14.

U.S. Department of Labor, *Monthly Labor Review*, various issues.

Accounting Rates of Return: Note

By Robert N. Anthony*

The excellent article on accounting rates of return by Franklin Fisher and the late John McGowan (1983) has reawakened interest in the inherent difference in the way profitability is measured by economists and by accountants. The comments and reply published in the June 1984 issue of this *Review* are evidence of this interest (Ira Horowitz, William Long and David Ravenscraft; Stephen Martin; Michael van Breda; Fisher). A substantial part of this difference between the measurement of profitability by accountants and by economists can be eliminated by making two changes in accounting principles. If economists join those academic accountants who advocate these changes, progress in reconciling the two disciplines is feasible.

The economists' task is to estimate a proposed project's profitability by comparing the amount to be invested in the project with the present value of the stream of estimated inflows *over the whole life of the project*. The accountant must report income *in each year* of the project's life. In order to do this, the accountant must chop the stream of inflows and the investment outflow into annual pieces. Most accounting problems (not only those relating to investments) are associated with the necessity of constructing annual income statements.

The difference between the economist's and the accountant's concept of income can be illustrated by a simple example. Assume that a company is considering an investment of $100,000 in a new process that is expected to produce cash inflows of $30,000 a year for five years (either in profits from additional sales or in cost savings). If the company has a required earnings rate of 15 percent, the economist calculates that the project would be profitable. The present value of the five annual inflows, discounted at 15 percent, is $100,565, which slightly exceeds the outflow of $100,000 at time zero.

If the proposal is accepted, the accountant's task is to report its income each year. Assuming that the actual outflow and the inflows are as expected, the accountant reports revenue of $30,000 a year. From this, the accountant subtracts depreciation expense, in order to match the investment with this annual revenue. The straightline method is used by the great majority of companies, and this method gives an annual depreciation expense of $20,000. The reported income is therefore $10,000 in each of the five years.

Table 1 shows that the return on investment (*ROI*) resulting from this calculation is inconsistent with the economist's correct calculation. We know that the *ROI* is slightly more than 15 percent in total. Since the inflows in each year are equal, the *ROI* is also 15 percent in each of the five years. The accounting calculation never results in such an annual *ROI*, whether the income of $10,000 is related to the beginning investment, the average investment, or the ending investment. In the first year the reported return on the beginning investment is less than 15 percent. The percentage increases each year thereafter, becoming ridiculously high in the later years. It does not average 15 percent. In no individual year is it 15 percent. More generally, the accounting calculation always understates the true return in the early years.

The discrepancy illustrated in Table 1 would be even greater if an accelerated depreciation method were used. It would also be greater for an investment whose cash inflows were relatively low in the early years; this is often the case with an investment for a new product or an expansion of capacity.

Two changes in accounting would correct this discrepancy: 1) charge annuity deprecia-

*Ross Graham Walker Professor of Management Control, Emeritus, Harvard Business School. I am grateful for comments of Charles J. Christenson and Franklin M. Fisher.

TABLE 1—CONVENTIONAL CALCULATION OF *ROI*

Year	Income	Investment		Return on Investment on[a]		
		Beginning	Ending	Beginning	Average	Ending
1	$10,000	$100,000	$80,000	10.0	11.1	12.5
2	10,000	80,000	60,000	12.5	14.2	16.7
3	10,000	60,000	40,000	16.7	20.0	25.0
4	10,000	40,000	20,000	25.0	33.3	50.0
5	10,000	20,000	0	50.0	100.0	

[a]Shown in percent.

TABLE 2—PROPOSED CALCULATION OF INCOME

Year	Beginning Investment	Revenue	Interest	Depreciation	Income
1	$100,000	$30,000	$15,000	$14,832	$168
2	85,168	30,000	12,775	17,056	168
3	68,112	30,000	10,217	19,615	168
4	48,497	30,000	7,275	22,557	168
5	25,940	30,000	3,891	25,940	168
		$150,000	$49,158	$100,000	$840

tion, and 2) recognize interest on capital—both debt capital and equity capital—as a cost. Neither idea is new. Annuity depreciation (often called "economic depreciation") was discussed in nineteenth-century literature. Its calculation is the same as that used to divide periodic loan amortization payments into principal and interest components. Most elementary economics texts state that the use of equity capital has a cost. Nevertheless, accountants do not recognize this cost in calculating a company's income.

Table 2 shows the income statements that would result from these two changes. The interest expense for a year is calculated as 15 percent of the amount of investment still outstanding at the beginning of that year. The depreciation schedule is set so that (a) depreciation plus interest is a constant amount in each year, and (b) the total depreciation equals $100,000. Interest expense decreases each year as a portion of the investment is recouped, and depreciation expense increases by a corresponding amount.

The reported annual income, $168, is the same each year, which is consistent with the facts. The amount is positive, indicating that the return is slightly more than 15 percent, also consistent with the facts. The net present value of the income stream is $565, the same as the amount calculated by the economist.

Table 2 was constructed on the assumption that the inflows occur on the last day of each year, consistent with the assumption usually made in present value calculations. An assumption that inflows occur earlier and more frequently would not change the point of the illustration.

As Hector Anton (1956) has shown, the depreciation calculation can be adapted to other flow patterns. The general approach is to arrive at an amortization schedule that 1) recovers the amount of the investment over the life of the project, and 2) charges interest on the unrecovered amount in a given year.

Accounting standards are established by the Financial Accounting Standards Board (FASB). If the FASB decides that companies should use annuity depreciation and account for equity interest, all companies that publish financial statements will be required to do so. The FASB has not made a pronouncement either for or against annuity depreciation. It is aware of the intellectual case for

recognizing equity interest, but it is troubled by three arguments that have been made against this principle.

First, recognizing equity interest as a cost would reduce the reported income of all companies. Conceptually, a company whose expenses, including the cost of using capital, exactly equalled its revenues would report zero net income. Economists already accept this idea. Accountants and users of financial statements will require a period of adjusting to the idea that zero income represents satisfactory performance.

Second, equity interest is regarded as an "imputed cost," and accounting is said not to recognize imputed costs. This is a fallacious argument; accounting does recognize costs that are not evidenced by documented transactions in events such as pensions, capital leases, zero-discount bonds, and sales at other than arm's length.

Third, it is argued that there is no feasible way of calculating the cost of using equity capital. This argument is weak. Although there is no precise way of calculating equity interest cost directly, the pre-tax interest rate for debt capital is a good approximation of the actual rate for total capital, that is, debt plus equity. The pre-tax debt rate overstates the actual debt cost because debt interest is tax deductible. It understates the equity cost because equity is riskier than debt. These two errors tend to balance each other out. I have shown (1983, ch. 4) that within a reasonable set of debt-equity ratios and of equity risk premiums, the pre-tax debt rate is within two percentage points of the true interest rate on total capital.

Although the arguments against charging annuity depreciation and recognizing the cost of using equity capital are weak, the task of persuading the FASB to adopt them is difficult. These proposals involve the most significant changes in accounting since the recognition of depreciation. The FASB will not adopt them unless there are strong pressures to do so. Such pressures do not now exist. If economists, either individually or through the American Economic Association (which has a formal channel to the FASB), convince the accounting profession that these changes will lead to accounting reports that conform more closely to economic reality, we just might see the FASB take some action.

REFERENCES

Anthony, Robert N., *Tell It Like It Was, A Conceptual Framework for Financial Accounting*, Homewood: Richard D. Irwin, 1983.

Anton, Hector R., "Depreciation, Cost Allocation, and Investment Decisions," *Accounting Research*, April 1956, 117–34.

Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

______, "The Misuse of Accounting Rates of Return: Reply," *American Economic Review*, June 1984, *74*, 509–17.

Horowitz, Ira, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 492–93.

Long, William F. and Ravenscraft, David J., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 494–500.

Martin, Stephen, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 501–06.

van Breda, Michael F., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 507–08.

The Inverted Fisher Hypothesis: Additional Evidence

By MARTIN GALLAGHER*

In their 1983 article, Jeffrey Carmichael and Peter Stebbing proposed the novel "inverted Fisher hypothesis." With respect to a major corollary of this hypothesis, they conclude their econometric analysis: "...according to these estimates not only is inflation not reflected in the after-tax nominal interest rate one-for-one, it is not reflected at all" (p. 624). The purpose of this note is to clarify this conclusion. Since "reflected" is not a well-defined statistical term, I was prompted to attempt to infer its meaning from the econometric evidence which supports the Carmichael and Stebbing (C-S) conclusion. In Section I, I show that this evidence is consistent with both a weak and a strong form of "unreflectedness." It is consistent with the contemporaneous uncorrelatedness of inflation and the net interest rate. It is also consistent with the stronger property of *both* contemporaneous and intertemporal uncorrelatedness. In Section II, an empirical analysis is presented that distinguishes between these cases.[1]

I. Two Forms of "Unreflectedness"

Carmichael and Stebbing present the inverted Fisher hypothesis as

$$(1) \qquad i_{Nt} = \alpha_0 + \xi_t,$$

where i_{N_t} is the net of tax nominal interest rate at time t, α_0 is a constant and ξ_t is a zero mean homoscedastic random variable.[2]

In order to test this hypothesis, Carmichael and Stebbing derive the estimating equation

$$(2) \qquad r_{Nt} = \alpha_0 + \alpha_2'\pi_t + \xi_t - (1+\alpha_2')\varepsilon_t,$$

where r_N is the after-tax real interest rate, π is the inflation rate, and ε is a mean zero homoscedastic random variable defined by

$$(3) \qquad \pi_t = E(\pi_t) + \varepsilon_t.$$

The inverted Fisher hypothesis (equation (1) above) may be tested by noting the identity

$$(4) \qquad r_{Nt} + \pi_t = i_{Nt}.$$

This implies that (1) is equivalent to the hypothesis that $\alpha_2' = -1$ in (2).

Carmichael and Stebbing point out that any errors in variables problem in (2) arising from (3) vanish under the null hypothesis that $\alpha_2' = -1$. They thus test the hypothesis by ordinary least squares (*OLS*) estimation. The hypothesis is accepted, and based on (1) it is concluded that π is "not reflected" in i_N.

This use of an *OLS* testing procedure involves a maintained hypothesis about the joint distribution of the sequences $\{\pi_t\}$ and $\{\xi_t\}$. The nature of this maintained hypothesis has implications for what may be construed about the relatedness of π and i_N from the acceptance of (1).

If the maintained hypothesis is that when $\alpha_2' = -1$ π is strictly exogenous in equation (2):

$$(5) \qquad E(\xi_t|\pi_s) = 0\ \forall t, s$$

the acceptance of (1) implies the acceptance of the contemporaneous *and* intertemporal uncorrelatedness of π and i_N. This may be seen by taking the expectation of (1) conditional on π_s and noting (5).

Alternatively, if the maintained hypothesis is that π and ξ are only contemporaneously uncorrelated $E(\xi_t\pi_t) = 0$, then, since from (1),

$$\mathrm{Cov}(i_{Nt}, \pi_t) = E(\xi_t\pi_t),$$

the acceptance of (1) implies only the contemporaneous uncorrelatedness of π and i_N.

*Department of Economics, University of Melbourne, Parkville, Victoria 3052, Australia.

[1]All estimation within this note involves the quarterly U.S. data for the period 1953:I–1978:IV described in detail by Carmichael and Stebbing. My thanks to Peter Stebbing for his provision of this data.

[2]No time subscripts appear in the original paper.

TABLE 1

Independent Variable	Dependent Variable: Rate of Inflation ($\Delta\pi_t$) (1)	Rate of Interest (Δi_{NT}) (2)
$\Delta\pi_t$		0.0302
		(1.06)
$\Delta\pi_{t-1}$	−0.8499	0.1082
	(−7.07)	(3.16)
$\Delta\pi_{t-2}$	−0.6882	0.1269
	(−4.44)	(3.48)
$\Delta\pi_{t-3}$	−0.4438	0.1118
	(−2.76)	(3.33)
$\Delta\pi_{t-4}$	−0.1680	0.0496
	(−1.13)	(1.81)
$\Delta\pi_{t-5}$	0.0626	
	(0.4768)	
$\Delta\pi_{t-6}$	0.0315	
	(0.26)	
$\Delta\pi_{t-7}$	0.0285	
	(0.25)	
$\Delta\pi_{t-8}$	−0.3558	
	(−3.07)	
$\Delta\pi_{t-9}$	−0.4755	
	(−3.90)	
$\Delta\pi_{t-10}$	−0.5203	
	(−3.90)	
$\Delta\pi_{t-11}$	−0.1622	
	(−1.20)	
$\Delta\pi_{t-12}$	−0.0008	
	(−0.007)	
Δi_{Nt}	0.5596	
	(1.43)	
Δi_{Nt-1}	0.8876	0.1501
	(2.28)	(−1.38)
Δi_{Nt-2}	1.2631	−0.3301
	(3.08)	(−2.99)
Δi_{Nt-3}	1.1269	−0.0851
	(2.73)	(−0.75)
Δi_{Nt-4}	−0.3013	−0.0203
	(−0.72)	(−0.19)
Constant	0.0005	0.00008
	(1.12)	(0.67)
SSE	0.00113	0.00012
Sample Size (T)	92	100

TABLE 2—LAGS FOR TABLE 1 ESTIMATIONS

lag(k) h_k	(1)	(2)
1	0.0858	0.5471
2	1.8739	0.5804
3	−0.5098	−0.6750
4	−0.8559	−1.2111
5	−1.3737	−2.0845
6	−0.6503	−1.8807
7	1.4581	−1.4231
8	0.1354	−0.4835
9	−0.0574	0.3724
10	−0.6987	−0.7256
11	1.8795	1.3653
12	0.4089	−0.7434

From this discussion it appears that the C-S evidence (i.e., the acceptance of (1)) does not permit us to distinguish between a strong form of both contemporaneous and intertemporal uncorrelatedness between π and i_N, and a weaker contemporaneous uncorrelatedness. I next present empirical evidence which distinguishes between these possibilities. Since the analysis of contemporaneous and intertemporal correlatedness is exactly what is addressed within Granger causality analysis, the analysis is conducted within a Granger causal framework.

II. Granger Causality Analysis of the Relationship Between π and i_N

Both π and i_N display obvious trend. Thus in the interests of stationarity, and to avoid the spurious regression cautioned against by C. W. J. Granger and P. Newbold (1974), the analysis was conducted in terms of Δi_N and $\Delta\pi$, where Δ is the first difference operator. Table 1 gives the results of *OLS* estimations of an equation specified to test the hypothesis that π is *not* Granger caused by i_N; (col. 1), and an equation specified to test the hypothesis that i_N is *not* Granger caused by π (col. 2).

In Table 1, *SSE* is the sum of squared errors from the regression, t-statistics are shown in parentheses and T is the sample size. In Table 2, h_k is the h-statistic at lag k for testing disturbance correlation in a model with lagged dependent variables. Under the null hypothesis of white noise disturbances, h_k is distributed as a standard normal variable.

Note that all h-statistics are insignificantly different from zero with the exception of h_5 in the rate of interest equation. In view, however, of the fact that a lag of 5 is not one at which correlation might be expected a priori to arise, and that conventional test size for the *joint* null hypothesis of white noise

residuals implies that a "small" test size is appropriate for the single test at each lag, I accept the null hypothesis of white noise residuals.[3]

To test whether or not i_N "Granger causes" π, I test the constraint that the coefficients on $\Delta i_{N(t-i)}$ in column 1 are zero for $i=0$ to $i=4$. Rejection of this constraint suggests that i_N "causes" π. To test the constraint, I use the Granger-Wald test recommended by John Geweke et al. (1983). Under the null hypothesis that the constraint is correct, the statistic

$$GW = T(SSE_c - SSE_u)/SSE_u$$

converges in distribution to a χ^2_c variate, where SSE_c and SSE_u are the sum of squared errors from the constrained regression and unconstrained regression, respectively, and c is the number of coefficients constrained to zero. In this case, $SSE_c = 0.00146721$ and $GW = 27.08$. Comparing this result to $\chi^2_{5,.01} = 15.1$, I reject the constraint. This result, allied with the insignificance of the coefficient on Δi_{Nt} prompts the conclusion that i_N Granger causes π, but *not* instantaneously.[4]

To test whether or not π causes i_N, I test the constraint that the coefficients on $\Delta\pi_{t-i}$ in Table 1, column 2, are zero for $i=0$ to $i=4$. Using the notation and methodology described above, I found $SSE_c = 0.000139987$ and $GW = 17.45$. Once again comparing this result to $\chi^2_{5,.01} = 15.1$, I reject the constraint. Allied with the insignificance of the coefficient on the contemporaneous term $\Delta\pi_t$, this result prompts the conclusion that π Granger causes i_N, but *not* instantaneously. (See Table 2 for the lags.)

To summarize the results of this section; the data supports the hypothesis that π and i_N are contemporaneously uncorrelated, it also supports the hypothesis that there is Granger "feedback" between π and i_N.

[3]A substantial h value at lag 5 suggests respecifying the rate of interest equation to include longer lags for Δi_N. I extended the maximum lag on Δi_N to 8. This in no way changed the subsequent results; however, the associated h correlogram suggested that the extended lag constituted a misspecification.

[4]Charles Nelson and G. William Schwert (1977) provide evidence that the *pre-tax* nominal interest Granger causes inflation. See also Schwert (1979; Section II).

III. Conclusion

Carmichael and Stebbing present evidence which prompts them to conclude that inflation π is not "reflected" in the net nominal interest rate i_N. I have observed that reflected is not a well-defined term and that the C-S evidence is consistent with both a weak and a strong form of "unreflectedness." Clarification of the C-S conclusion is thus required.

Granger causality analysis has clarified the situation. The data supports the hypothesis that π and i_N are contemporaneously uncorrelated. The data does *not* support the hypothesis that π and i_N are intertemporally uncorrelated.

REFERENCES

Carmichael, Jeffrey and Stebbing, Peter, "Fisher's Paradox and the Theory of Interest," *American Economic Review*, September 1983, *73*, 619–30.

Geweke, John, Meese, Richard and Dent, Warren, "Comparing Alternative Tests of Causality in Temporal Systems," *Journal of Econometrics*, February 1983, *21*, 161–94.

Granger, C. W. J. and Newbold, P., "Spurious Regressions in Econometrics," *Journal of Econometrics*, July 1974, *2*, 111–20.

Nelson, Charles and Schwert, G. William, "Short-Term Interest Rates as Predictions of Inflation: On Testing the Hypothesis that the Real Rate of Interest is Constant," *American Economic Review*, June 1977, *67*, 478–83.

Schwert, G. William, "Tests of Causality, The Message in the Innovations," in Karl Brunner and Allan Meltzer, eds., *Three Aspects of Policy and Policymaking: Knowledge, Data, and Institutions*, Vol. 10, Carnegie-Rochester Conferences on Public Policy, *Journal of Economic Literature*, Suppl. 1979, 179–86.

Rationing by Waiting Lists: An Implication

By JOHN G. CULLIS AND PHILIP R. JONES*

In a recent article, Cotton Lindsay and Bernard Feigenbaum (1984) present and test a model of rationing by waiting lists. Its novel feature is the recognition that being on some types of waiting list involves no opportunity cost and that consumers' surplus cannot be dissipated by waiters undertaking costly activities that will help secure the good or service in question. In this sense, time does not act as a price although it imposes costs. Although Lindsay's earlier version (1980) of this model has already been misinterpreted by some commentators,[1] its heart is a waiting list that is equilibrated by attacking the assumption that the demand curve remains unchanged throughout the wait. Waiting time matters because the value of the good or service decays the longer it is delivered after order day. While not wishing to take issue with this insight, there are a number of points that need to be borne in mind when assessing the significance of the model, especially in relation to the authors' application to Britain's National Health Service (NHS).

I. Welfare Costs and Waiting Lists

Although the numbers recorded on in-patient waiting lists are high and have risen since the inception of Britain's NHS in 1948, interpretation is a difficult matter. Almost by definition, waiting lists are for conditions that are capable of waiting. This produces a very biased concentration of waiting lists (a feature consistent with Lindsay and Feigenbaum). Waiting lists relate mainly to certain surgical specialties and are of minor importance for medical specialties. Crude waiting list figures are unreliable and include individuals whose admission has been delayed for domestic or medical reasons, as well as some who no longer require in-patient treatment.

Some indication of the actual composition of lists can be gained from a recent review of an orthopedic waiting list. The review revealed some surprising evidence concerning those waiting for more than a year to be called as in-patients at a large district general hospital (L. J. Donaldson et al., 1984). Of the 950 potential patients involved in the review, a preliminary record search revealed 20 percent had already been treated and, of the remaining 757, a postal questionnaire indicated that only 48 percent still wanted the operation. The others had either died[2] (5 percent), moved away (9 percent), already received treatment (9 percent), or were nonrespondents (12 percent). Interestingly, the remaining 17 percent no longer wanted operations. There was evidence that longer waiting times increased the proportion no longer seeking treatment. The major reasons for this were either that minor conditions proved self-correcting, or, in the major operations category, that patients were too ill or frail to proceed.

The importance of this in the context of Lindsay and Feigenbaum is twofold. First, waiting list data have to be viewed with caution and, on this evidence, they considerably overstate the position. Second, even if the NHS waiting list (when adjusted for the

*School of Humanities and Social Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, England.

[1] "Lindsay develops a rather elegant but simple model of waiting time based on supply and demand analysis. His characterization of the rationing process is, however, somewhat misleading for it implies that some patients are deterred from joining the list as a result of the costs associated with waiting. Since the chief waiting costs such as anxiety, pain, and inconvenience must be endured whether or not the patient joins the list, it is difficult to see how such costs would deter patients from joining. While Lindsay's model, therefore, may be appropriate for explaining the demand for general practitioners' services where patients may incur time costs in waiting rooms, an alternative explanation of the rationing process for hospital services is required."

[M. W. Spicer, 1982, p. 667]

[2] All were over age 65.

considerations outlined above) remains "one of the largest queues in the Western world" (Lindsay-Feigenbaum, p. 405) there still exists the question of its welfare interpretation. Lindsay and Feigenbaum suggest relatively little about the welfare costs of rationing by waiting lists, but the implication is that they are very large. In fact, an application of their analysis makes possible an estimate of the costs of waiting and the result achieved would call into question the significance of this feature of the NHS.

The analysis presented by Lindsay and Feigenbaum permits a taxonomy of potential consumers of medical treatment and a broad indication of the private costs each category experiences as a result of waiting lists. Below V_0 is the individual valuation of treatment now, C the cost of joining the waiting list, and t the waiting time on the list. To proceed, it is important to be precise as to the nature and relevance of C. In the NHS system, these are likely to be dominated by the initial costs of examination, diagnosis, and referral by a general practitioner. It is at this stage of referral to a consultant that the patient is placed on the hospital waiting list.

Figure 1 illustrates the position of marginal waiters, individuals who may be considered to be indifferent as between going private now (t_0) or joining the queue to receive care at $\hat{t}$. Let us begin by considering the individual on boundary A in Figure 1. (The case is one to which Lindsay and Feigenbaum refer.) If waiting time is $\hat{t}$, the individual will accept the initial costs C of joining the waiting list provided that g (the decay rate) is no greater than g_0. It would not be worth the individual accepting these costs C if waiting time is $\hat{t}$ and $g > g_0$. If $g = g_0$, it is clear that the present value of treatment at $\hat{t}$ is just equal to C so that the net present value of joining the waiting list to the individual is zero. It is the case then that there is a minimum estimate of the value of treatment for marginal waiters.

The individual at the stage of referral, however, has another alternative. By purchasing *private treatment* at a price P he may avoid waiting.[3] In this way, while there

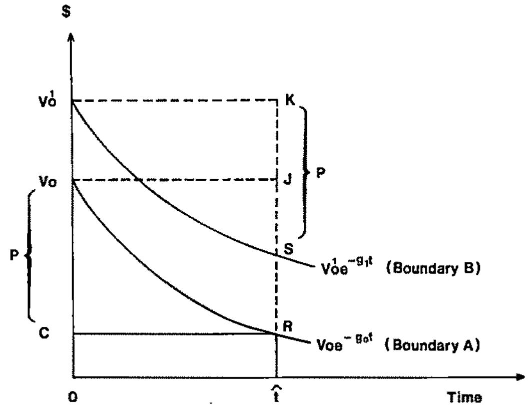


FIGURE 1. MARGINAL WAITERS

is a minimum estimate to the value of joining a waiting list, there is also a maximum limit to the costs of waiting. In Figure 1 the individual on boundary A is deemed to have a value of treatment now (i.e., $[(V_0 - C) - P]$) of zero. Therefore he is a marginal waiting patient in that there is no differential advantage in securing private treatment now over and above waiting (i.e., $[(V_0 - C) - P] = (V_0 e^{-g_0 \hat{t}} - C)$). While those with $(V_0 - C) = P$ but $g > g_0$ will not wait, those with $(V_0 - C) = P$ but $g < g_0$ will be intramarginal waiters in the sense that there is a positive inducement to join the waiting list even though $(V_0 - C) = P$. A subset of waiters (with intercept $P = (V_0 - C)$ on the y axis) includes all those with $g \leq g_0$. For the individual on boundary A, the cost of waiting is clearly P, but for others in this subset, it must be less than P. The maximum value of the costs of waiting, P, can only arise if the benefits from treatment are completely dissipated by the wait and here the positive net present value of treatment after waiting $(V_0 e^{-g\hat{t}} - C) > 0$ is a mitigating factor. Since this subset then includes those with $g = g_0$ to $g = 0$ (i.e., falling within V_0JR in Figure 1) the cost of waiting may range between P and zero.

This same perspective may be applied to other subsets for whom $(V_0 - C) > P$. Consider the marginal waiter on boundary B in Figure 1. He is a person for whom $(V_0^1 - C)$ signifies the benefit of immediate treatment.

[3] See Private Patients Plan (1984).

While for this individual $(V_0^1e^{-g_1\hat{t}}-C)$ is positive, the issue as to whether or not he will wait is resolved by the question of whether or not $(V_0^1e^{-g_1\hat{t}}-C)\geq[(V_0^1-C)-P]$. In Figure 1, for the individual on boundary B, there is no differential advantage in going private or waiting (i.e., $(V_0^1e^{-g_1\hat{t}}-C)=[(V_0^1-C)-P]$). Clearly, however, g may now exceed a specific decay rate g_1 and yet $(V_0^1e^{-g\hat{t}}-C)$ remain positive but the individual still will not wait since this would imply that $(V_0^1e^{-g\hat{t}}-C)<[(V_0^1-C)-P]$. Thus a decay rate, g_1, is again critical in the sense of equating the net benefit of immediate treatment with that of waiting. Those with $g\leq g_1$ wait. For these, once again, the maximum cost of waiting (on boundary B) is P and for those with $g<g_1$ (falling in area V_0^1KS in Figure 1) the maximum dissipated benefit from waiting ($KS=P$ in Figure 1) will not be experienced.

For this subset the range of dissipated benefit from waiting can never exceed P, but may lie between zero and P. It follows that the same is true for all individuals who have a valuation (on the y axis) such that $(V_0-C)>P$; that is, dissipated benefits from waiting can never exceed P but may fall to zero.

This analysis leaves us with another category of waiters which we refer to as submarginal. These are individuals who pass the criteria for waiting (i.e., $V_0e^{-g\hat{t}}\geq C$), but for whom $(V_0-C)<P$. They would not pay for treatment but, as it is provided at no personal price, they will opt to wait. Therefore the dissipated benefit can never equal P for these submarginal waiters. Their net valuation of treatment (V_0-C) may be positive, but is insufficient to outweigh the costs of treatment. They present a problem of moral hazard for the NHS but, in private terms, *they* can hardly be construed as bearing costs of waiting. If treatment within the NHS did not exist, they would not purchase treatment and their decision to wait reflects the benefit of this alternative. Without the NHS waiting list, (V_0-C), which may be positive, would dissipate but, given the NHS, $(V_0e^{-g\hat{t}}-C)$ may be positive and hence this is a benefit for them. We regard this category as likely to be few in number within the NHS waiting lists though their existence reflects private benefit rather than cost and in this sense make calculations below an overestimate.

TABLE 1—TYPICAL COSTS OF 1984 U.K. PRIVATE TREATMENT

Operation/Treatment	From	To
Appendectomy	£610	£1160
Hernia Operation	610	1160
Hysterectomy	1000	2200
Duodenal Ulcer	949	1750
Cataract	880	1250
Tonsillectomy	440	880
Slipped Disc Repair	1500	2900
Gall Bladder Removal	1060	1650
Varicose Veins	560	1060
Bunion Removal	750	1250
Knee Cartilage	1000	1900
Heart Operation	3500	5000
Breast Lump Removal	500	940
Hip Replacement	1600	3100
Prostate Removal	1000	2150
Nasal Polyp Removal	400	940
Prolapse Repair	1000	2200
Bowel Resection	1500	2250

Source: Private Hospital Plan.

Following Lindsay-Feigenbaum and regarding g as dominated by the decay rate (rather than time preference), we conclude that the costs of waiting (i.e., the dissipated benefits), can never exceed P. It is in this way that the Lindsay-Feigenbaum analysis throws light on the private costs imposed on those waiting in the NHS. As accepted by Lindsay-Feigenbaum, waiters will be those for complaints which do not have a high decay rate. Table 1 outlines the range of prices for private treatment of such complaints. The range is quite high for any given treatment because of variations in consultant fees and hospital costs.[4] However, the striking impact of these prices is the fact that they provide an upper price estimate that can be multiplied by the waiting list in the NHS to get a maximum value of the private costs of waiters. To allow for the fact that P is an upper bound of the net benefit loss and that the range of the waiting costs for indi-

[4]Such hospital charges are outlined by the Department of Health and Social Security (DHSS) (1983).

viduals can, a priori, be expected to be randomly distributed from zero or near zero to the upper bound, the average value of dissipated benefit loss is assumed to be $\frac{1}{2}P$.

Given this assumption, half the average prices (lower and upper) from Table 1 are taken as typical and the mean waiting time for all conditions for England and Wales for 1981 given in the Hospital In-Patient Enquiry (DHSS, 1984) is combined with an overall U.K. waiting list in 1981 of approximately 745,000 (Central Statistical Office (CSO), 1985a) to calculate a "ballpark" welfare cost for waiters. With a mean waiting time of 16.8 weeks, the implied number of waiting patients treated in the NHS per annum will be 2.3 million, that is, by implication, the waiting list turns over three times per annum approximately. Multiplying by half the upper average (£937) gives £2,155.1 million and for the lower price average (£524) £1,205.2 million. These are large absolute sums, and for each individual affected the cost of waiting may be high, as indicated in Table 1. However, as percentages of government expenditure on the NHS in 1981–82 (i.e., £13,267 million, CSO, 1985b), they would be 16.2 and 9.1 percent, respectively. In 1982, using mean waiting time for England only, the same calculation yields 18.8 and 10.5 percent, respectively. But what of the others who "go private" or "balk"?

The question then arises as to how those who go private and those who neither go private nor wait are affected by the waiting list. For the former, it will depend on their expectations as taxpayers. Realistically, payment of taxes simply purchases the property right of access to a waiting list and, to this extent, they have no welfare loss at time t_0 when they additionally pay privately for treatment. On this assumption the crude calculations above therefore need no further amendment.

The remaining group of individuals who balk at both the waiting time and the private option are likely to be small since such potential patients must have simultaneously high decay rates so that $V_0e^{-g\hat{t}} < C$ and potential net benefit from treatment now $(V_0 - C) < P$. What may be of more significance (as Lindsay-Feigenbaum suggest) are those with both $(V_0 - C) > P$ and high g's who are treated immediately in the NHS but not, as ideally might be the case, as in-patients. It is difficult to know how many patients are involved. It can be noted, however, that the associated (private) efficiency calculation involves the *change* in output valuation without in-patient treatment but with the actual treatment regime adopted. If these costs were large, there would surely be more direct evidence of it than has surfaced in Britain, where acute care appears, at minimum, satisfactory and is in fact a source of pride in the medical profession (see Royal Commission, 1979). If these last arguments are valid, the crude welfare cost calculation survives and, while not wishing to "go over the top" on its strength, it may nevertheless be indicative of the figure that would emerge from a more sophisticated analysis.

That the private welfare costs of waiting lists may not be unduly high finds other support, for example, in the Royal Commission on the National Health Service. The Commission reported that "Surprisingly, waiting lists attracted little comment in our evidence" (p. 126) and noted that the OPCS survey found that 80 percent of all in-patients said they were not caused inconvenience or distress by waiting for admission.

Recently a Marplan survey was based on a quota sample of 1500 adults, randomly selected from a sampling frame that controlled for sex, age, social class, and geographical location, in order to test public opinion of the NHS (S. Halpern, 1985). Results indicated that only 12 percent of the public had a bad opinion of the NHS while for those who had received treatment, 62 percent were "very satisfied," 25 percent were "fairly satisfied," and only 13 percent had a neutral opinion or were dissatisfied. Out of those dissatisfied with their treatment, 22 percent gave delays in receiving an appointment or treatment as the reasons. However, in the context of apparent general satisfaction, this hardly appears damning.

Unfortunately, all that has been said so far must be qualified if it is to form the basis of public policy decisions. The analysis has been based completely upon the private decision-making calculus of Lindsay and Feigen-

baum's model, where the individual decides whether or not to join the waiting list if the initial costs C are equal to the future discounted value of treatment. It is argued that, as the delay is considered in terms of treatment, g is determined primarily by the decay rate component rather than the time preference discount rate component.[5] Thus the joining criterion is $C = V_0/(1+g)^t$ and the waiting list is the outcome of such decisions.

It is, however, questionable that this criterion should form the basis of waiting list determination. It is clear, for example, that such private decision making completely ignores the (future) costs of treatment because they are not priced to the individual. This discussion corresponds to the submarginal waiters above who are induced to wait even though $(V_0 - C) < P$.[6] Their treatment would be inefficient and such cases can be expected to be discouraged in the NHS which attempts to solve this moral hazard problem by rationing care according to "need."

It is clear then that the costs of treatment cannot be ignored and therefore management of the waiting list may be determined by broader considerations than those captured in Lindsay and Feigenbaum's model. The "optimal" waiting list that arises from such management need not equal zero. To accept this would imply that the NHS stand ready to meet at any time any demand placed upon it. Thus while costs of a managed waiting list need not equal zero, they will not necessarily mean that the waiting list should be cut unless it is also proved that the associated benefits exceed costs at the margin. If one agreed with the normative ground rules of a social efficiency framework, employed appropriately to manage the waiting list, then, while waiting lists and waiting costs are positive, it would be incorrect to perceive this outcome as basis for criticism.

There are a number of different criteria that may determine the waiting list. They depend upon the normative framework of who should decide upon admission to the list and upon what basis. The following are not exhaustive but are illustrative of different options society may employ.

Private calculus as described by Lindsay and Feigenbaum would imply a minimum joining criterion of $C = V_0/(1+g)^t$. The lump sum (transactions) costs now to the individual of joining the list being equated with the individual perception of benefit discounted at g over the expected mean waiting time.

Paretian social calculus would imply a joining criterion whereby the social discounted costs of joining the list are equated with the social discounted benefits. Benefits would be reduced by the decay rate, but now both sides of the equation may also be influenced by a social discount rate. The benefits will be more broad; for example, social benefits of adding a person to the waiting list may include the better balance of intake for major and minor operations and for teaching purposes. Cost considerations may include the benefit implicit in pushing operation costs into the future, net of interim treatment costs. However, the final cost of operation and treatment is a cost which society must consider. The context is clearly broader than the Lindsay-Feigenbaum framework.

Social welfare calculus would imply even further amendment to the joining criteria. For example, valuation of treatment may be amended by a weight which accords to the opinion of "experts" as to the "need" or "deservingness" of special cases who join the list.[7] Such a decision may be influenced by income distribution and equity more broadly defined.

The implication of such considerations are twofold. First, it is impossible to dogmatically argue that the waiting lists and their attendant welfare costs are either "too small" or "excessive" without explicit reference to a normative benchmark. Second, it may well be the case that waiting lists and waiting costs are positive even if the solution is deemed optimal. Having said this, the

[5] Though the value of treatment for all patients decays at the same rate, the time preference will differ between consumers.

[6] The implicit assumption here is that P represents the marginal cost of treatment both in the NHS and the private sector.

[7] See, for example, A. J. Culyer and Cullis (1976), and C. E. B. Frost (1980) for a discussion of criteria.

Lindsay-Feigenbaum contribution is still valuable, because it does permit the estimation of some boundaries upon the costs to patients who wait within the NHS. Yet it is only in the light of the above qualifications that any normative interpretation can be applied to the estimates presented. As always, there is scope for argument.

II. Data and Result Matters

Lindsay and Feigenbaum recognize that their data based on fourteen NHS regions are not ideal, in part because waiting lists are a hospital rather than a regional concern and waiting times vary both between and within regions. The authors argue, however, that patients will, to a great extent, equalize waits within regions by "shopping around," so that their regression results are not significantly distorted. Unfortunately this conjecture finds little support. As they point out, being placed on a list is the result of a G.P. referral and a consultant outpatient appointment. Shopping around is not easily practiced in this framework and the data confirm this assertion. For example, there are marked contrasts in regional waiting times for orthopedic surgery but the differences among health districts are even greater (DHSS, 1981). The Royal Commission (p. 127) cite further evidence: in 1978, South Camden had 73 people per thousand on the waiting list, while North Camden had less than one-seventh of this; in central Manchester, there were 39 per thousand waiting compared to 6 in neighboring Bury. To be fair to Lindsay and Feigenbaum, this evidence relates to waiting lists rather than waiting times which are observations in their empirical work. However, there is evidence that waiting lists and waiting times are positively related.

Taking a cross section of the fourteen English hospital regions for 1978, and marrying the mean waiting time data for all diagnostic categories (DHSS, 1981) with regional waiting list data for all specialties (Office of Health Economics, 1984) via a linear regression of mean waiting times on a constant and waiting lists, yields a positive coefficient of 0.08 on waiting lists with a t-statistic of 2.39. More significant for intraregional issues is the recent report of the College of Health which has as its first main conclusion "...that if you live in any of the districts with a long wait for certain specialties then you should usually be able to find a hospital which has a short wait in another district *within* the same region" (1985, p. 24, emphasis added). Lindsay and Feigenbaum have worked with the data to hand but, while accepting this data limitation, there is reason to be somewhat cautious of their findings and to await a more disaggregated analysis.

Such data limitations notwithstanding, it is interesting to consider Lindsay and Feigenbaum's discussion of the effect of increasing capacity on the waiting list. They argue that "...so long as the elasticity of joining with respect to expected wait has an absolute value greater than one, an increase in supply will result in a waiting list containing *more* names!" (p. 411). This observation can be placed against the background of the debate about the "agency" role of consultants[8] in waiting list formation and the resistance of lists to increased consultant supply. The absolute values of the estimated elasticities for the low decay-rate conditions ranged from 0.55 to 0.64 and for high decay-rate conditions from 0.65 to 0.7. Other things equal, this tends to suggest that waiting lists should respond favorably to increasing consultant supply. The Lindsay-Feigenbaum results therefore appear to leave any observations of increased supply and waiting lists without an easy interpretation and do not reject the agency hypothesis.

III. Conclusion

The discussion above suggests two major concerns in relation to Lindsay and Feigenbaum's article. First, that while the welfare costs of rationing are not a simple matter, a broad indication of their significance can be established on the basis of individual valuations. The existence of an "exit-voice" facility permits such estimation. Here a starting point has been offered into such exploration. At this stage, however, the evidence

[8]For a discussion of this topic see Frost (1977, 1980, 1981), Frost and B. J. Francis (1979) and K. McPherson (1980, 1981).

suggests that while such costs are significant, they are not overwhelming. It is difficult to overestimate the significance of this implication of the Lindsay-Feigenbaum contribution. Since the advent of the NHS, waiting lists have been an emotive point of discussion among academics, politicians, and the media. Often they are used as a "big stick" with which to belabor the NHS, within a debate more notable for political rhetoric than for measured evidence. To the extent that the Lindsay-Feigenbaum framework proves an acceptable analysis, certain upper and lower bounds to the costs of waiting must be implied. Of course, such valuations are in terms of willingness to pay and a question mark may be raised on the issue of income distribution, but this is a matter for broader public policy measures than simple waiting list management.

Second, attention has been drawn to the empirical work of Lindsay and Feigenbaum. Here it is impossible to agree that certain problems concerning the limitations of the data can be discounted. While sympathetic to the constraints under which the analysis was presented, the results must be treated tentatively until a more disaggregated study can be furnished. Finally, with this caveat firmly in mind, it is worth stressing that empirical estimates for the "joining" elasticity actually do imply the existence of other explanations for the response of waiting lists to increased supply. Other factors must influence waiting lists and the agency hypothesis is not easily dismissed.

REFERENCES

Culyer, A. J. and Cullis, J. G., "Some Economics of Hospital Waiting Lists in the NHS," *Journal of Social Policy*, July 1976, *5*, 239–64.

Donaldson, L. J., Maratos, J. I. and Richardson, R. A., "Review of an Orthopaedic In-Patient Waiting List," *Health Trends*, February 1984, *16*, 14–5.

Frost, C. E. B., "Clinical Decision Making and the Utilization of Medical Resources," *Social Science and Medicine*, 1977, *11*, 793–9.

______, "How Permanent are NHS Waiting Lists?," *Social Science and Medicine*, 1980, *14C*, 1–11.

______, "Clinical Decision Making: A Reply," *Social Science and Medicine*, 1981, *15C*, 43–45.

______ **and Francis, B. J.,** "Clinical Decision Making: A Study of General Surgery within Trent RHA," *Social Science and Medicine*, 1979, *13A*, 193–98.

Halpern, S., "What the Public Thinks of the NHS," *Health and Social Service Journal*, June 6, 1985, *94*, 702–04.

Lindsay, Cotton M., *National Health Issues: The British Experience*. Nutley: Roche Laboratories, 1980.

______ **and Feigenbaum, Bernard,** "Rationing by Waiting Lists," *American Economic Review*, June 1984, *74*, 405–17.

McPherson, K., "Clinical Decision Making: C. E. B. Frost and B. J. Francis," *Social Science and Medicine*, 1980, *14C*, 285.

______, "Clinical Decision Making: A Response to a Reply," *Social Science and Medicine*, 1981, *15C*, 193–96.

Spicer, M. W., "The Economics of Bureaucracy and the British National Health Service," *Milbank Memorial Fund Quarterly*, Winter 1982, *6*, 657–72.

Central Statistical Office, (1985a) *Social Trends 15*, London: HMSO, 1985.

______, (1985b) *Annual Abstract of Statistics*, No. 121, London: HMSO, 1985.

College of Health, *Guide to Hospital Waiting Lists in England and Wales*, 2d ed., London: College of Health, 1985.

Department of Health and Social Security (DHSS), *Orthopaedic Services: Waiting Time for Out-Patient Appointments and In-Patient Treatment; Report of a Working Party to the Secretary of State for Social Services*, London: HMSO, 1981.

______, *Health Services Management Charges for Private Patients*, Health Circular, HC (83)7, March 1983.

______, *Hospital In-Patient Enquiry 1981*, London: HMSO, 1984.

Office of Health Economics, *Compendium of Health Statistics*, 5th ed., London: Office of Health Economics, 1984.

Private Patients Plan, *Private Hospital Plan*, Private Patients Plan Ltd., April 1984.

Royal Commission on the National Health Service, (Chairman, Sir Richard Merrison), Cmnd. 7615, London: HMSO, July 1979.

Implicit Contracts in the Absence of Enforcement: Note

By STEVEN E. PLAUT*

In recent years, implicit contract theory has grown following the work by Martin Baily (1974) and Costas Azariadis (1975). A major concern in much of this literature has been the problem of enforceability of implicit contracts. The problem is that for any contract to trade labor services, or some other good at some future time t, there will always be motivation for one of the contracting parties to breach the contract whenever the future spot price at t deviates from the contractual price. In the absence of some formal enforcement mechanism (i.e., courts), or informal mechanism (such as concern for reputation or front-end loading), contracts would never be fulfilled.

In an important contribution to this literature, Clive Bull (1983) proposed a model under which implicit contracts would become partially enforceable due to what might be called a "package contract" for two distinct labor services, which he called labor and effort. Bull argued that the inability to trade effort separately due to the nonexistence of a market for such would lead to implicit contracts that were partially enforceable (i.e., in at least some states of the world). His conclusion was that "the very aspect of the economy that gives rise to implicit contracts [is]...the absence of a complete set of markets" (p. 668).

It is the theme of this note that the absence of markets for some aspects of labor is unnecessary in order to achieve the Bull results of partial enforceability. While such enforceability is possible when some labor markets are missing, it is no less possible when labor markets are complete (although contingent claims markets must remain incomplete).

To see this, let us follow Bull and imagine that a worker is supplying two distinct labor services, which will be called "work" and "effort." Each of these may be marketed separately to different purchasers, or the two may be marketed jointly in a multiple contract.[1]

Now the future prices of work (or X_t) and of effort (or Y_t) are unknown before t, and no contingent claims or enforceable forward markets for them exist. In the absence of formal legal enforcement or informal enforcement (such as concern for reputation), no contract for supplying either labor service separately would be fulfilled. A contract to supply work at price $\hat{X}_t$ would be broken by workers as soon as the future spot price rose above that, and would be breached by employers as soon as it fell below $\hat{X}_t$. The same would hold for separate contracts to purchase effort.

But what about joint or multiple contracts to supply both work and effort simultaneously? Let the indirect utility function of the worker be $\beta_t U(X_t, Y_t)$, where β_t is the discount factor for future period t. Let the indirect utility function of the employer or purchaser be $\beta_t W(-X_t, -Y_t)$. For both the U and W functions, the first partials with respect to both arguments are positive and the second partials are negative. Both U and W are continuous, differentiable functions and the Inada conditions hold.

Let us assume there is a contract to jointly trade work and effort at promised prices of

*Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Technion City, Haifa, 32 000, Israel.

[1]It is not difficult to show examples of labor contracts involving two or more separate labor services that could be conceivably marketed independently. A receptionist/typist could do typing for firm B in between answering phones at firm A. For an example closer to home, an economist could teach courses at institution A while doing research and publishing papers under the auspices of institution B, being paid according to his outputs separately and independently. While intrafirm economies of scope could explain the existence of the joint contracts we commonly observe, so could the existence of voluntarily enforceable multiple contracts. In the absence of formal enforcement mechanisms, only such multiple contracts would be viable.

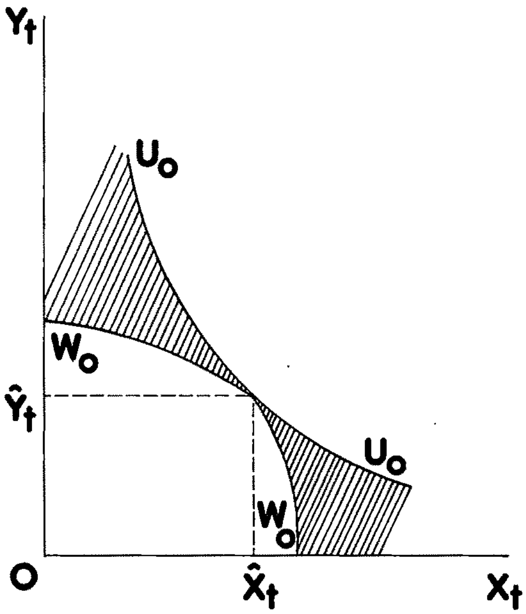


FIGURE 1

$\hat{X}_t$ and $\hat{Y}_t$, respectively, at future date t. For simultaneous trade that really takes place at these prices, the worker will receive utility U_0 and the employer will receive W_0.[2] (See Figure 1.) It is noted that any future spot price combination (X_t, Y_t) that falls within the shaded area will result in the contract being voluntarily enforced. The reason is that while each party would gain by breaching one of the contract clauses for one type of labor service transaction as long as the other clauses were being enforced, neither party would have any motivation to breach the contract if this would result in the nonfulfillment of *both* contractual transactions.[3]

In sum, the lack of separate markets for the distinct labor services traded in the market is not a necessary condition for partially enforceable implicit labor contracts. The ability of contracting parties to enter into joint implicit agreements for more than one labor service is sufficient to produce voluntary enforcement and therefore may explain the existence of implicit contracts in the absence of enforcement mechanisms.

[2] The implied quantities derived from differentiating the indirect utility functions are assumed to be the same for both parties, allowing for market "clearing." In Figure 1, the contractual prices are Pareto optimal. Actually even Pareto nonoptimal contract price combinations will be partially enforceable (i.e., in some states of the world) for similar reasons and could lead to partially enforceable implicit contracts.

[3] Moreover, since there is some probability that the contract will be enforced voluntarily at t (namely the probability that future prices will be in the shaded area), the contract itself would have value to both parties even before t. It could then serve to reinforce voluntary compliance with contractual obligations even before t if noncompliance resulted in nullifying later contractual commitments.

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin Neil, "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Bull, Clive, "Implicit Contracts in the Absence of Enforcement and Risk Aversion," *American Economic Review*, September 1983, *73*, 658–71.

Efficient Contracts in Credit Markets Subject to Interest Rate Risk: An Application of Raviv's Insurance Model

By LANNY ARVAN AND JAN K. BRUECKNER*

Interest rate risk borne by financial institutions has increased markedly as the credit markets have become more unstable in recent years. The burden of this new instability has been especially severe for lending institutions whose balance sheets show the greatest mismatch between the maturities of assets and liabilities. The savings and loan industry, for example, which borrows short and lends long, suffered serious losses in recent years as income from its portfolio of old mortgages failed to keep pace with the escalating cost of short-term funds. The specter of such losses has led to the emergence of adjustable-rate mortgages (*ARMs*) as well as other types of variable-rate contracts. These contracts often specify a complex functional relationship between the future loan interest rate and the lender's (uncertain) future cost of funds.

In view of the growing popularity of variable-rate loans, it is important to know what features an efficient contract of this type would possess. To answer this question, the present paper derives the form of the optimal variable-rate contract, focusing especially on the contractual relationship between the loan rate and the future cost of funds. The framework used in the analysis is very similar to Artur Raviv's (1979) optimal insurance model. Section I derives the form of the optimal loan contract and evaluates the risk-sharing efficiency of *ARMs* in light of the results. Subsequent sections discuss extensions of the model and offer conclusions.[1]

*Assistant Professor and Professor of Economics, respectively, University of Illinois at Urbana-Champaign, Urbana, Il 61801. We thank Case Sprenkle, James Follain, and a referee for comments. Any errors are ours.

[1] Several earlier papers deal with the maturity mismatch problem that underlies the present study; see Sudhakar Deshmukh et al. (1983a, b) and Jürg Niehans and John Hewson (1976).

I. Analysis

The model has a single borrower and a single lender who extends a two-period loan of exogenous size L. The borrower pays interest on the loan at the end of each period, with the principal repaid at the end of the second period. The lender is assumed to rely on short-term borrowing for his loanable funds, which means the terms of the loan contract must be set before the cost of funds in the uncertain second period is known. As a result, the contract specifies a loan-rate function $r(\cdot)$, which gives the second-period loan interest rate as a function of the short-term rate s that actually prevails in that period. The lender's net income in the second period is thus $(r(s)-s)L$, with net income in the first period equal to $(r_0-s_0)L$ (s_0 is the first period's short-term rate, which is freely observable, while r_0 is the loan rate). Letting y_0 and y denote the borrower's incomes in the two periods, net incomes are y_0-r_0L and $y-r(s)L$ in the first and second periods, respectively (y is nonrandom).[2] Inflation is assumed to be absent, so that all net incomes are in real terms.[3]

[2] Although the manner in which the borrower uses the loan is not relevant to the analysis, one possibility is that the funds are spent directly on consumption, in which case $y_0=w_0+L$ and $y=w-L$, where w_0 and w denote exogenous incomes. Alternatively, the proceeds of the loan could be used to purchase a capital good that generates incomes in the two periods according to the relationships $y_0=h_0(L)$ and $y=h(L)$. Finally, the loan could finance purchase of a consumer durable such as a house. While utility in this case will depend on the services from the durable (as measured by L) as well as on net income, the service argument of the utility function is suppressed under the above formulation (note that in the latter two cases, the loan principal is repaid with the proceeds from the sale of the asset).

[3] The analysis is essentially unchanged when inflation is introduced provided that price increases are nonrandom. Stochastic inflation, however, changes the character of the results.

The loan contract just described can be viewed as a type of insurance arrangement, as in Raviv. The lender's interest payment s to depositors plays the role of the random loss x in Raviv's insurance model. The borrower's payment $r(s)$ covers this loss in the same manner that the insurer's payment of $I(x)$ in Raviv's model helps protect the insured party. The parallel between the models is not exact, however, since the loan model lacks an explicit analog to the insurance premium P,[4] and has no equivalent to the insurer's administrative cost $c(I(x))$.

Letting v denote the lender's utility function (which satisfies $v'' \leq 0$), $\delta < 1$ denote his discount factor, and $f(s)$ denote the density function for s, the lender's discounted expected utility equals

$$(1)\quad v[(r_0 - s_0)L] + \delta \int_0^{\bar{s}} v[(r(s) - s)L] f(s)\, ds,$$

($\bar{s}$ is the maximum value of s). The optimal loan contract (which specifies r_0 as well as the loan-rate function $r(\cdot)$) is chosen to maximize (1) subject to the requirement that the borrower's discounted expected utility equals some constant. Letting u denote the borrower's concave utility function and $\theta < 1$ denote his discount factor, this constraint is written

$$(2)\quad u(y_0 - r_0 L) + \theta \int_0^{\bar{s}} u(y - r(s)L) f(s)\, ds = k,$$

where k is a constant. The maximization problem is solved by choosing the loan-rate function optimally conditional on r_0 and then optimizing over r_0. Under this procedure, $r(\cdot)$ is chosen to maximize the expected utility integral in (1) subject to the requirement that the expected utility integral in (2) equals the constant $(k - u(y_0 - r_0 L))/\theta$.

Crucial constraints in Raviv's model are the requirements that the insurance payment be nonnegative and no larger than the loss $(0 \leq x \leq I(x))$. While no constraints are explicitly imposed on the loan-rate function in deriving the main results of the analysis, assumptions leading to an upper bound on $r(\cdot)$ are discussed below.

The Hamiltonian for the problem of choosing $r(\cdot)$ is $\{v[(r(s)-s)L] + \lambda u(y - r(s)L)\} f(s)$, where λ is the costate variable. The optimality conditions are $\dot{\lambda} \equiv d\lambda/ds = 0$ and

$$(3)\quad v'[(r(s) - s)L] - \lambda u'(y - r(s)L) = 0,$$

which show that the optimal loan-rate function equates the lender's marginal utility to a constant proportion of the borrower's marginal utility regardless of the realized value of s. Using (3), it can be shown that optimal choice of r_0 yields a standard intertemporal efficiency condition.[5] The slope of the loan-rate function is found by differentiating (3) with respect to s, which gives $(\dot{r} - 1)v'' + \lambda \dot{r} u'' = 0$. Eliminating λ using (3) and solving yields

$$(4)\quad \dot{r} = \sigma_v / (\sigma_u + \sigma_v),$$

where $\sigma_v \equiv -v''/v'$ and $\sigma_u \equiv -u''/u'$ are the absolute risk-aversion measures for the lender and borrower, respectively (the σ's are evaluated at the relevant net income levels). Raviv's equation (10), which gives the slope of the I function when $0 < I(x) < x$, reduces to (4) when the insurer's administrative cost is independent of the size of the payment.

Equation (4) expresses a simple and intuitively appealing rule for the sharing of interest rate risk in credit markets. The equation shows that when both parties are risk averse,

[4] Borrowers receive a payment for risk absorption in the form of a lower first-period interest payment or a lower loan-rate function.

[5] The condition is

$$\delta \int_0^{\bar{s}} v' f\, ds / v'_0 = \theta \int_0^{\bar{s}} u' f\, ds / u'_0,$$

where v'_0 and u'_0 represent marginal utilities in the first period. This condition states that the expected marginal rates of substitution between net incomes in the two periods must be the same for lender and borrower.

the slope of the optimal loan-rate function is positive and lies between zero and one, indicating that a percentage point change in the lender's cost of funds yields less than a percentage point change in the loan rate r. As a result, lender and borrower share the burden of a higher cost of funds as well as the benefit of a lower cost. More importantly, however, (4) shows that the manner in which risk is allocated between lender and borrower depends on the relation between absolute risk-aversion measures. The equation shows that if the borrower is locally more (less) risk averse than the lender, $\dot{r}(s)$ should be less than (greater than) 1/2, indicating that the borrower locally bears a smaller (greater) share of the interest rate risk than the lender. Note that while the borrower should bear no interest rate risk at all when the lender is risk neutral ($\dot{r}=0$ when $\sigma_v=0$), it is optimal for the borrower to absorb all the risk if he himself is risk neutral ($\dot{r}=1$ when $\sigma_u=0$). Although the slope of the loan-rate function in general varies with s, it is constant when both lender and borrower exhibit constant absolute risk aversion. In this familiar case, the loan-rate function is a straight line.

While the above analysis applies to any credit market where the maturities of assets and liabilities are mismatched, the results are especially useful in evaluating the growing tendency for risk sharing in today's mortgage market. Prior to the 1980's, the fixed-rate mortgage (which satisfies $\dot{r}=0$) was the mainstay of the savings and loan industry. The heightened volatility of interest rates in recent years, however, inflicted large losses on the industry and led to introduction of the *ARM*, in which the borrower absorbs interest rate risk in return for a lower interest cost. The most common *ARM*s have the following structure. Within a prespecified band (as shown in Figure 1), the loan rate is set equal to the short-term rate (or some cost-of-funds index) plus a markup at each adjustment period. Within the band, *ARM*s satisfy $\dot{r}=1$ and imply complete risk absorption by the borrower. Interest rate caps determine the location of the band and prevent the loan rate from rising or falling excessively between adjustment periods (see Figure 1).[6] When a cap is binding, $\dot{r}=0$ and the lender absorbs all interest rate risk.[7]

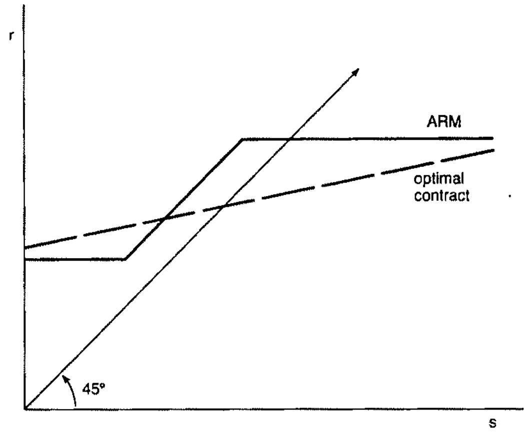


FIGURE 1

Inspection of Figure 1 shows that *ARM*s offer at best a crude approximation to the optimal risk-sharing rule embodied in (4). Before concluding that existing contracts are inefficient, however, it is important to ask whether a modified model might yield a different verdict. Two related alterations are in fact capable of generating an upward cap on r as part of an optimal contract. The first is a bankruptcy constraint on the borrower, which is written $y-r(s)L\geq 0$ and implies an upward cap at y/L. Another approach is to incorporate bankruptcy of the insured party (the lender) into the model. Gur Huberman et al. (1983) show that under this modification, the optimal insurance contract has a maximum payment, again implying an upward cap on r in the present context.[8] While both these approaches are suggestive, neither can explain the presence of down-

[6] The caps are symmetrical around the previous period's loan rate.

[7] See Brueckner (1985) for an analysis of consumer choice in the *ARM* market (the choice between adjustable- and fixed-rate mortgages is the main focus).

[8] The idea is that as a result of limited liability, the insured may prefer to lower his premium by choosing a policy with a maximum payment. When a large loss occurs, the insured limits liability by choosing bankruptcy.

ward caps in existing contracts.[9] Furthermore, since risk sharing is still optimal when the cap is nonbinding under both modifications, the $\dot{r}=1$ requirement of existing contracts remains inexplicable.[10]

II. Extensions

Here we consider two extensions of the model. Under the first extension, the lender is allowed to choose his loan volume by adjusting the number of borrowers served (the loan size remains fixed for simplicity). Letting n denote the number of borrowers and replacing L in (1) by nL, the first-order condition for choice of n is

$$(5)\quad (r_0 - s_0)v_0' + \delta\int_0^{\bar{s}}(r(s)-s)v'f(s)\,ds = 0,$$

where v_0' denotes the lender's marginal utility in the first period. Repeating the previous analysis, the slope of the loan-rate function becomes

$$(6)\qquad \dot{r} = n\sigma_v/(\sigma_u + n\sigma_v).$$

While the appearance of the endogenous variable n on the right-hand side of (6) complicates the connection between $\dot{r}$ and the risk-aversion measures, the effect of a change in σ_v is easily derived when absolute risk aversion is constant. In this case, it may be shown that $n\sigma_v$ (and hence $\dot{r}$) is invariant to a change in σ_v, with n falling as σ_v rises and vice versa.[11] Since the intercept of the (linear) loan-rate function and the value of r_0 are also invariant to a change in σ_v, it follows that the optimal loan contract is independent of the risk aversion of the lender. The only effect of an increase in risk aversion is a reduction of the lender's optimal loan volume. This result is intuitively plausible given that the variability of the lender's net income $(r(s)-s)nL$ can be reduced in response to a higher σ_v either by an increase in $\dot{r}$ holding n fixed, or by a reduction in n holding the loan-rate function fixed.

As a second extension of the model, the borrower's second-period income is taken to be random and correlated with the lender's cost of funds.[12] Let $F(s, y)$ denote the joint density of these variables, with $G(y|s)$ denoting y's conditional density. Then the borrower's expected utility in the second period becomes $\int_0^{\bar{s}}\int_0^{\bar{y}}u(y-r(s)L)F(s,y)\,ds\,dy$, and the Hamiltonian for the control problem (with $n=1$) is rewritten as[13]

$$(7)\quad \left\{v[(r(s)-s)L] + \lambda\int_0^{\bar{y}}u(y-r(s)L)G(y|s)\,dy\right\}f(s).$$

With constant absolute risk aversion, the slope of the loan-rate function is given by

$$(8)\qquad \dot{r} = (\sigma_v - \Omega)/(\sigma_u + \sigma_v),$$

where

$$(9)\qquad \Omega = \int_0^{y}u'\frac{\partial G}{\partial s}\,dy \Big/ L\int_0^{y}u'G\,dy.$$

In the case where y and s are jointly normal, it may be shown that $\Omega \gtrless 0$ as $\rho \lessgtr 0$, where ρ

[9] It should be noted that Raviv's nonnegativity constraint on I is a type of downward cap imposed via the natural but somewhat arbitrary assumption that payments must flow only from the insurer to the insured. Given its arbitrariness, the analogous assumption $r(s) \geq 0$ is not imposed in the present model. Moreover, it seems impossible to justify the assumption required to generate observed downward caps ($r(s) \geq b > 0$).

[10] While $\dot{r}=1$ is optimal if the borrower is risk neutral, this does not appear to be a realistic possibility.

[11] This can be seen by writing equations (2), (5), and the condition from fn. 4 for the constant absolute risk-aversion (exponential utility) case using a linear r with slope (6). In the equations, n and σ_v only appear in the product expression $n\sigma_v$, establishing that if $\hat{n}$ solves the optimization problem with $\sigma_v = \hat{\sigma}_v$, then $\tilde{n} = \hat{n}\hat{\sigma}_v/\tilde{\sigma}_v$ solves the problem when $\sigma_v = \tilde{\sigma}_v$.

[12] The incomes of workers employed in industries whose fortunes are especially sensitive to interest rate movements (such as housing) will presumably be negatively correlated with s.

[13] If y were freely observable to the lender, then the optimal contract would make the loan rate contingent on both y and s, with a loan-rate function of the form $r(s, y)$. For the above formulation to represent an optimum, y must be unobservable.

is the correlation coefficient.[14] In this case, (8) and (9) imply that if the borrower's income is positively (negatively) correlated with the lender's cost of funds, it is optimal for him to bear more (less) interest rate risk than when y and s are independent, a natural result.

III. Conclusion

This paper has analyzed the optimal variable-rate loan contract in a framework similar to Raviv's optimal insurance model. In addition to showing that a variable-rate contract can be viewed as a type of insurance arrangement, the paper derived results of current practical interest. In particular, the optimal risk-sharing rule emerging from the model was shown to provide a useful starting point for evaluating the efficiency of existing variable-rate contracts. This is an important contribution since the variety of such contracts can be expected to multiply in today's increasingly innovative financial environment.

[14] With normality, the numerator of Ω is proportional to $\rho\int_0^{\bar{y}} u'(y - E(y|s))G(y|s)\,dy$. Since u' is a decreasing function of y and the remainder of the integrand integrates to zero, it follows that the integral multiplying ρ is negative.

REFERENCES

Brueckner, Jan K., "The Pricing of Interest Rate Caps and Consumer Choice in the Market for Adjustable-Rate Mortgages," unpublished paper, 1985.

Deshmukh, Sudhakar D., Greenbaum, Stuart I. and Kanatas, George, (1983a) "Interest Rate Uncertainty and the Financial Intermediary's Choice of Exposure," *Journal of Finance*, March 1983, *39*, 141–47.

______, ______, **and** ______, (1983b) "Lending Policies of Financial Intermediaries Facing Credit and Funding Risk," *Journal of Finance*, June 1983, *38*, 873–86.

Huberman, Gur, Mayers, David and Smith, Clifford W., "Optimal Insurance Policy Indemnity Schedules," *Bell Journal of Economics*, Autumn 1983, *14*, 415–26.

Niehans, Jürg and Hewson, John, "The Eurodollar Market and Monetary Theory," *Journal of Money, Credit and Banking*, February 1976, *8*, 1–27.

Raviv, Artur, "The Design of an Optimal Insurance Policy," *American Economic Review*, March 1979, *69*, 84–96.

Faculty Ratings of Major Economics Departments by Citations: An Extension

By DUDLEY W. BLAIR, REX L. COTTLE, AND MYLES S. WALLACE*

Paul Davis and Gustav Papanek (1984) ranked major economics departments by citations; however, their approach was not novel. Dennis Gerretz and Richard McKenzie (1978) initiated the use of citations to rank economics departments. Though they ranked only southern economics departments for the years 1976 and 1977, Gerritz-McKenzie used both total citations and mean citations as ranking criteria.

As Davis-Papanek and Gerritz-McKenzie have shown, the citation-ranking approach eliminates many of the problems of ranking departments using journal articles and offers a qualitative as well as quantitative measure of faculty productivity. Both studies, however, ranked only Ph.D. granting institutions and ignored those economics departments which offer a master's degree as the highest degree awarded. Although many Ph.D.-granting departments have master's programs, these institutions tend to concentrate on training future academicians and to treat the master's degree as a consolation prize for unsuccessful Ph.D. students.

While non-Ph.D.-granting institutions may have different incentives, teaching loads, computer and library support, and quality and extent of graduate research assistantships, we use a citations-based criterion to demonstrate that these institutions should not be neglected as sources of economic research.[1]

Column (1) of Table 1 shows the ranking of economics departments offering the master's degree as the highest degree based upon the average number of citations for the years 1977–1981 for department faculties denoted in the 1982 catalogs of their respective institutions (ties are ranked equally). Our procedure differs from Davis-Papanek in that they averaged citations from 1978 and 1981 only.

Given the wide range of faculty size, column (2) denotes the mean citation per faculty member per year. As Davis-Papanek found for Ph.D.-granting institutions, the rankings do change significantly. Four of the top ten departments are replaced by smaller departments and there is considerable shuffling among the remaining top ten departments.

Since Davis-Papanek included faculty members citing their own work, this may bias one of the major advantages of using citations instead of the number of major journal publications; that is, citations measure the quality of a person's research in stimulating further research by others. Column (3) ranks departments by citations per year adjusting for self-citations. In some cases this correction is important. For example, Auburn moves from 9th to 15th, Brigham Young from 5th to 13th, and West Texas State from 12th to 6th. In general, there is a high correlation between the two rankings (the Spearman rank correlation coefficient is .99).

Another potential weakness of a citation index is the fact that one article cited ten times is weighted equally with ten articles cited once each. It would be an interesting exercise to examine how innovations in the literature (as measured by citations per article) compare between master's only institutions, and middle- to lower-ranked Ph.D. programs. Because Davis-Papanek do not separate the number of articles cited from the number of citations, we are unable to make this comparison. For this reason, the number of articles is not reported in our tables. For master's only institutions, the

*Blair and Wallace: Department of Economics, Clemson University, 222 Sirrin Hall, Clemson, SC 29631; Cottle: University of Mississippi. We thank an anonymous referee for helpful suggestions on a earlier draft. Any remaining errors are our own.

[1]Philip Graves et al. (1982) included master's only institutions in their publications-based rankings.

TABLE 1—UNIVERSITY RANKING OF INSTITUTIONS GRANTING MASTER'S DEGREES BY TOTAL, PER CAPITA, AND NON-SELF PER CAPITA CITATIONS (1977–81 Annual Averages)

Rank (1)	University or College	Total Citations Number	Citations per Capita Rank (2)	Citations per Capita Number	Citations per Capita Excluding Self-Citations Rank (3)	Citations per Capita Excluding Self-Citations Number
1	CUNY-Baruch	104.0	3	3.59	3	3.46
2	CUNY-Queens	87.4	1	5.14	1	4.75
3	Delaware	54.0	11	2.35	9	2.22
4	Nevada-Reno	53.4	2	4.11	2	3.69
5	Cal State-Long Beach	50.6	6	2.53	5	2.44
6	Auburn	47.6	9	2.38	15	1.91
7	CUNY-City	34.6	4	3.15	4	3.05
8	Cleveland State	32.4	14	2.16	10	2.15
9	So. Illinois-Edwards	31.8	13	2.27	12	2.06
10	Akron	28.0	15	2.00	13	1.94
11	CUNY-Hunter	26.2	8	2.38	8	2.31
12	San Francisco State	25.8	16	1.98	16	1.91
13	Denver	24.2	7	2.42	11	2.12
14	San Jose State	23.6	19	1.69	19	1.54
15	Clemson	21.2	27	1.18	28	0.98
16	Cal State-Northridge	20.2	32	1.01	30	0.96
17	Brigham Young	19.8	5	2.83	13	1.94
17	CUNY-Brooklyn	19.8	28	1.16	25	1.06
19	New Mexico State	19.0	51	0.61	48	0.61
20	Cal State-Fullerton	18.0	34	0.95	35	0.83
21	Drake	16.6	10	2.37	7	2.31
22	Hartford	16.0	24	1.23	22	1.18
22	Wichita State	16.0	31	1.07	26	1.04
24	Miami (Ohio)	15.4	37	0.81	36	0.80
24	Texas-Arlington	15.4	29	1.10	29	0.97
26	Butler	15.2	18	1.90	17	1.88
27	Bowling Green State	14.2	40	0.79	43	0.71
28	West Texas State	14.0	12	2.33	6	2.33
29	East Texas	13.6	17	1.94	18	1.86
30	Canisius	13.4	22	1.34	21	1.30
30	San Diego State	13.4	54	0.58	52	0.57
30	Western Michigan	13.4	33	0.96	32	0.91
33	Illinois Inst. Tech	13.2	20	1.47	20	1.42
33	Toledo	13.2	29	1.10	27	1.02
35	Rhode Island	13.0	26	1.18	23	1.18
36	Portland State	12.4	42	0.77	40	0.75
37	NC-Greensboro	11.8	46	0.74	49	0.60
38	Duquesne	11.6	23	1.29	24	1.07
38	South Florida	11.6	59	0.53	58	0.48
40	Cal State-Chico	10.8	25	1.20	31	0.96
41	New Orleans	10.6	58	0.53	56	0.51
41	Western Illinois	10.6	50	0.66	53	0.56
43	Illinois State	10.0	48	0.71	45	0.67
43	Old Dominion	10.0	44	0.77	42	0.72
45	Kent State	9.8	52	0.61	51	0.57
46	Cal State-LA	8.4	71	0.38	69	0.34
47	Nebraska-Omaha	8.0	38	0.80	41	0.74
47	Puerto Rico	8.0	47	0.73	47	0.65
49	Murray State	7.4	36	0.82	39	0.76
50	Sangamon State	7.2	53	0.60	50	0.58
51	No. Texas State	7.0	41	0.78	55	0.51
52	Orono, MA	6.8	45	0.76	44	0.69
53	Wright State	6.4	62	0.49	60	0.48
54	George Mason	6.2	68	0.41	73	0.29
54	Tulsa	6.2	42	0.77	38	0.77
54	SUNY-Oneonta	6.2	78	0.31	72	0.31

(continued)

TABLE 1—Continued

Rank (1)	University or College	Total Citations Number	Citations per Capita Rank (2)	Citations per Capita Number	Citations per Capita Excluding Self-Citations Rank (3)	Citations per Capita Excluding Self-Citations Number
57	Mankato State	6.0	64	0.46	71	0.32
58	Baylor	5.8	70	0.39	76	0.28
59	Bridgeport	5.6	56	0.56	54	0.56
59	Cal State-Fresno	5.6	56	0.56	59	0.48
61	Cal State-Hayward	5.4	66	0.45	63	0.45
62	Texas-El Paso	5.2	63	0.47	78	0.25
63	Marquette	5.0	74	0.36	75	0.29
63	East Tennessee State	5.0	61	0.50	57	0.50
63	De Paul	5.0	81	0.25	85	0.21
66	Miami	4.6	55	0.57	61	0.47
66	Virginia Commonwealth	4.6	85	0.24	92	0.19
68	East Illinois	4.4	79	0.29	79	0.25
69	Indiana State-Terre Haute	4.2	71	0.38	67	0.36
69	Texas Women's	4.2	49	0.70	46	0.67
71	Central Michigan	4.0	90	0.20	91	0.19
71	No. Dakota	4.0	73	0.36	68	0.35
71	Roosevelt	4.0	38	0.80	36	0.80
74	Bucknell	3.6	60	0.51	66	0.37
75	Long Island	3.4	35	0.85	34	0.85
76	Alaska	3.2	75	0.36	86	0.20
76	Colorado-Denver	3.2	65	0.46	62	0.46
78	Loyola	3.0	81	0.25	80	0.25
78	Cal State-Poly	3.0	81	0.25	84	0.22
80	Humboldt State	2.8	69	0.40	65	0.40
80	Atlanta	2.8	21	1.40	33	0.90
80	Montana	2.8	77	0.31	74	0.29
83	East Michigan	2.4	86	0.24	81	0.24
84	Jersey City	2.2	67	0.44	64	0.44
85	So. Dakota	2.0	76	0.33	70	0.33
86	Ball State	1.8	96	0.13	98	0.11
86	Seton Hall	1.8	94	0.15	100	0.12
88	Mississippi State	1.6	89	0.23	83	0.23
88	Idaho	1.6	80	0.27	77	0.27
90	Fairleigh	1.4	100	0.12	98	0.12
90	Western Kentucky	1.4	88	0.23	86	0.20
92	Montclair	1.2	86	0.24	81	0.24
92	Florida Central	1.2	97	0.12	95	0.12
94	Detroit	1.0	102	0.07	102	0.07
94	St. Mark	1.0	81	0.25	86	0.20
94	Texas Christian	1.0	93	0.17	93	0.17
97	Central Missouri State	0.8	101	0.10	101	0.10
97	Connecticut	0.8	90	0.20	86	0.20
99	Florida Atlantic	0.6	97	0.12	95	0.12
99	Hardin-Simmons	0.6	90	0.20	86	0.20
99	West Florida	0.6	97	0.12	95	0.12
102	No. Colorado	0.4	95	0.13	94	0.13
103	Fort Hays State	0.2	103	0.04	103	0.04
104	Gannon					
104	Mississippi State for Women					
104	Kansas State-Pitts.					
104	Goddard					
104	Inter American					
104	Missouri-Kansas City					
104	Sam Houston State					
104	St. Thomas					
104	Trinity					

Notes: Rankings based on: col. 1: average number of citations denoted in their respective 1982 catalogs; col. 2: mean citation per faculty member per year; col. 3: departments by citations, adjusting for self-citations.

TABLE 2—RANKING OF SELECTED INSTITUTIONS OFFERING NO GRADUATE DEGREES IN ECONOMICS, BY TOTAL, PER CAPITA, AND NON-SELF PER CAPITA CITATIONS (1977–1981 Annual Averages)

		Total Citations	Citations Per Capita		Citations per Capita Excluding Self-citations	
Rank (1)	School	Number	Rank (2)	Number	Rank (3)	Number
1	Wesleyan	154.4	3	12.87	2	11.98
2	Swarthmore	121.2	1	17.31	1	16.54
3	Brandeis	72.6	4	4.84	4	4.76
4	Colby	28.6	6	3.18	7	2.24
5	Sarah Lawrence	28.2	2	14.10	3	11.10
6	Amhearst	27.6	7	2.76	6	2.68
7	Smith	27.2	10	1.60	10	1.55
8	William & Mary	22.0	11	1.57	11	1.39
9	Wellesley	17.2	12	1.32	12	1.22
10	Oberlin	11.8	8	1.97	8	1.97
11	Reed	11.2	5	3.73	5	3.60
12	Middlebury	11.2	9	1.87	9	1.77
13	Vassar	5.8	13	0.72	13	0.72
14	Hamilton	1.0	14	0.13	14	0.10

Notes: See Table 1.

simple regression coefficient between mean citations and mean number of articles cited per faculty member is 1.34 ($R^2 = .93$). If an article is cited, it is cited only slightly more than once. When ranks are compared, the Spearman rank correlation is .99, indicating little difference between the rankings for master's only institutions.

In contrasting our rankings by total citations to those of Davis-Papanek (assuming our five-year and their two-year averages are comparable), the following comparisons can be made.

1) Seventy-six of the non-Ph.D., master's departments would rank above the 122nd Ph.D. department.

2) The top four non-Ph.D., master's departments would rank in the top 100 Ph.D. departments.

3) The top non-Ph.D. master's institution would rank 84th.

A handful of selected schools which offer no graduate degrees in economics were also examined. As reported in Table 2, the highest-ranked school by total citations, Wesleyan University, ranks above the highest ranked master's only institutions and would rank 73rd among Ph.D.-granting institutions. When citations per capita are used, the top school, Swarthmore College, would rank 33rd in the Ph.D. list, and the top three undergraduate only colleges would rank in the top forty-four institutions.

Even if compared on absolute levels, Ph.D.-granting departments do not monopolize the market for economic ideas. Master's only departments and some departments having only undergraduate programs have made significant contributions to the economic literature and should be included in any rankings of major economics departments.

REFERENCES

Davis, Paul and Papanek, Gustav F., "Faculty Ratings of Major Economics Departments by Citations," *American Economic Review*, March 1984, *74*, 225–30.

Gerritz, Dennis M., and McKenzie, Richard B., "The Ranking of Southern Economics Departments: New Criterion and Further Evidence," *Southern Economic Journal*, October 1978, *45*, 608–614.

Graves, Philip E., Marchand, James R. and Thompson, Randall, "Economics Departmental Rankings: Research Incentives, Constraints, and Efficiency" *American Economic Review*, December 1982, *72*, 1131–41.

Structural/Frictional vs. Deficient Demand Unemployment: Comment

By Arthur R. Schwartz, Malcolm S. Cohen, and Donald R. Grimes*

In a recent article in this *Review*, Katherine Abraham (1983) compared structural/frictional with demand deficient unemployment using two alternative measures of the unemployment-to-vacancy ratio. The first measure was based on vacancy rates from various surveys, while the second measure used information on the new hire rate to construct a vacancy rate. It is the second measure that will be discussed here.

Abraham assumes a steady state for vacancies; that is, the number of new vacancies flowing into the stock will equal the number of vacancies that are filled from the stock. Since the duration of a vacancy in the steady state is equal to the stock of vacancies divided by the flow, the duration can be written as

$$d_v = V/NH \tag{1}$$

where V = number of vacancies, NH = number of new hires, and d_v = duration of a vacancy.

If we divide numerator and denominator by employment and solve for the vacancy *rate*, we get

$$NHR \cdot d_v = VR, \tag{2}$$

where NHR is the new hire rate and VR is the vacancy rate.

Abraham is attempting to compare the "stock" of unemployed persons with the stock of vacancies. The shorter the duration of a given vacancy, the lower the steady-state stock of vacancies. If, for example, each vacancy were open for only one hour, then at any point in time, the stock of vacancies would be approximately zero.

Our research on labor turnover can be used to better understand the dynamics of the labor market. We will use our new hires research results to make four main points concerning the Abraham article.

1) The number of new hires in the private nonfarm economy is much greater than Abraham's data indicate.

2) Vacancies are rarely in a steady state as seasonal variation is extensive.

3) There are two classes of hires: permanent and temporary. Lumping these types of new hires together could give a distorted view of how the labor market is working.

4) The vacancy-unemployment ratio varies dramatically by region and industry. An aggregate rate could mask problems in the labor market.

The new hire rate that was used in Abraham understates total hiring activity because it measures persons and not transactions. She used the *CPS* job tenure question to compute the new hire rate. By using this question, she is measuring the number of *persons* who were new hires in January of a particular year. This understates the total number of vacancies that were available during that period, since a person could be a new hire on more than one occasion. Our research has shown that, over the course of a year, it is quite likely that many people will be new hires more than one time.

We have calculated the number of new hire *transactions* by using a 1 percent sample of Social Security Administration records to follow individuals over time. The methodology is documented in detail elsewhere, but essentially it consists of following unique employee Social Security and employer identification numbers from a firm's Social

*Assistant Research Scientist, Director, and Research Associate, respectively, Institute of Labor and Industrial Relations, University of Michigan, Victor Vaughan Bldg., 1111 East Catherine Street, Ann Arbor, MI 48109–2054.

Security report.[1] If an individual's Social Security number is not on an employer's report in any of the four previous quarters and appears in the current quarter, then that person is a new hire in the current quarter. This methodology provides a means to measure the total number of hiring transactions in the private, nonfarm economy.

The new hire rate for 1973, first quarter, from the Social Security wage records was 22.3 percent. This is a monthly rate of 7.4, assuming that hires are uniform from month to month over the quarter. This is considerably higher than the 4.7 used by Abraham. If we use Abraham's assumption that vacancies were open on the average of 10 days, then the comparable vacancy rate is 2.4. This compares to 1.5 used by Abraham. Table 1 compares the statistics used by Abraham to those that result from our recalculation. These figures exclude recalls since Abraham did not consider them. Our new hire rates are for wage and salary workers in the private nonfarm economy and thus the unemployment rate is also for that group of workers. Abraham's data presumably include agricultural and government workers.

These figures would imply that the unemployment-vacancy ratio for January 1973 is one-third lower than the number presented by Abraham. Even though our rate is lower than Abraham's rate, it is still high considering that the unemployment rate has never returned to the rate for 1973, which was 4.9 percent for the entire year. Also, many of the jobs that were vacant were very short-term jobs. In fact, only about one-third of the new hires were still with the same employer two quarters after the hire.[2]

It can be seen in Table 2 that the unemployment-vacancy ratio varies considerably with the cycle. In 1975 and 1982, very high unemployment years, the ratio was over five

TABLE 1

Variable	January 1973	
	Abraham	Social Security Records
Unemployment Rate	5.5	5.8
New Hire Rate[a]	4.7	7.4
Vacancy Rate[a]	1.5	2.4
Implied Unemployment-Vacancy Ratio	3.82	2.50

[a]Per 100 employees.

to one. Only in 1973 which was a very low unemployment year was the ratio near two to one.[3] It appears that unemployment rates greater than 5.8 percent are associated with unemployment-vacancy ratios greater than 2.5. Unemployment rates of 9 percent or greater appear to be associated with unemployment-vacancy ratios greater than five. Abraham has an unemployment-vacancy ratio of two with an unemployment rate of 4 percent and a ratio of seven with an unemployment rate of 7 percent.

Table 2 shows that the rates vary considerably over the cycle. New hire rates also have seasonal variation. Table 3 presents data for 1973 on a quarterly basis. The figures have been converted to monthly averages for comparability. As one can see from Table 3, there is some basis for questioning the steady-state assumption due to the seasonality of new hires.

The use of aggregate unemployment and vacancy data fails to allow for the fact that there are different kinds of hires. Of all the new hires in 1973, 65 percent were no longer in their new job two quarters after the hire. These temporary new hires are thus quickly back in the pool of job seekers. Only 35 percent of the hires can be classified as permanent.[4] It is important to differentiate these

[1]See Cohen and Schwartz (1980) for details on the construction of the new hire rates. Quarterly data from Social Security records are only available through 1977. However, quarterly data from state Unemployment Insurance records are available for many states on a current basis.

[2]See Cohen and Schwartz (1983).

[3]The new hire rates after 1975 are based on a forecasting model of new hire rates developed from the 1 percent Social Security sample. For details see our 1983 study. For all years it was assumed that the average duration of a vacancy was 10 days.

[4]We have defined a permanent new hire as a hire that remains on the job for at least two more quarters after the hiring quarter. This can be derived by following Social Security payroll data.

TABLE 2

Year	Private, Nonfarm Wage and Salary Workers: Unemployment Rate	Private, Nonfarm Wage and Salary Workers: Average Monthly New Hire Rate[a]	Average Monthly Vacancy Rate[a,b]	Unemployed-Vacancy Ratio
1973	4.9	7.7	2.5	2.01
1974	5.8	7.4	2.4	2.50
1975	9.2	5.9	1.9	5.23
1976	7.9	6.5	2.1	4.00
1977	7.0	6.8	2.2	3.35
1978	5.9	7.4	2.4	2.55
1979	5.8	7.3	2.4	2.50
1980	7.4	6.6	2.2	3.55
1981	7.7	6.4	2.1	3.89
1982	10.2	5.4	1.8	6.26

[a]Per 100 employees.
[b]Assuming 10-day duration.

TABLE 3

Quarter	Unemployment Rate	New Hire Rate[a]	Vacancy Rate[a]	Unemployed-Vacancy Ratio
1973:1	5.7	7.4	2.4	2.46
1973:2	4.7	8.1	2.6	1.85
1973:3	4.5	8.3	2.7	1.70
1973:4	4.4	6.9	2.2	2.05

[a]Per 100 employees.
[b]Assuming 10-day duration.

two types of hires in order to understand how the labor market works. The duration that a vacancy is open is a key variable in Abraham's analysis. The length of time the job remains filled is another key variable for labor market analysis. If the labor market provides a lot of short-time hires, then all we may have is a rotating pool of unemployed moving rapidly from unemployment to employment and back to unemployment. If a vacancy is filled for just a short period of time and then the employee is once again unemployed, should that vacancy be counted the same as a vacancy that keeps a worker employed for a long period of time? It is possible to have a worker counted as unemployed for an entire quarter, and yet fill a vacancy for three weeks (between *CPS* unemployment surveys). In this case, a vacancy would not have taken even one person out of the pool of unemployed. Many short-term new hires could cause the effective unemployment-to-vacancy (*U-V*) ratio to be too low. Vacancy data cannot measure the length of hire; turnover data can. It is important to study the length of a hire in order to formulate a comprehensive policy to deal with unemployment.

As the aggregate unemployment rate hides differences among industries and geographic regions, so too does the aggregate *U-V* ratio. Table 4 provides information by major industry class and Table 5 provides information for a sample of states. Both tables pro-

TABLE 4

Industry	Unemployment Rate	New Hire Rate[a]	Vacancy Rate[a,b]	Unemployed-Vacancy Ratio
Manufacturing	4.3	5.6	1.8	2.45
Services	4.8	7.7	2.5	1.97
Construction	8.9	14.6	4.8	1.94
Transportation and Public Utilities	3.0	4.9	1.6	1.90
Trade	5.6	9.7	3.2	1.79
Finance	2.7	5.7	1.9	1.43
Mining	2.8	6.3	2.1	1.34
Total	4.9	7.7	2.5	2.01

[a]Per 100 employees.
[b]Assuming 10-day duration.

TABLE 5

State	Unemployment Rate	New Hire Rate[a]	Vacancy Rate[a,b]	Unemployed-Vacancy Ratio
Michigan	5.9	6.6	2.2	2.79
New York	5.4	6.3	2.1	2.66
Pennsylvania	4.8	5.7	1.9	2.60
California	7.0	9.0	3.0	2.43
Ohio	4.3	6.4	2.1	2.09
So. Carolina	4.1	7.4	2.4	1.74
Arizona	5.0	11.0	3.6	1.41
Nevada	6.0	13.3	4.4	1.39
Georgia	3.9	9.2	3.0	1.31
Kansas	3.0	8.3	2.7	1.11
Florida	4.3	12.6	4.1	1.05
Nebraska	2.0	7.6	2.5	0.80

[a]Per 100 employees.
[b]Assuming 10-day duration.

vide annual data for 1973 and both use the assumption that the average duration for a vacancy is 10 days.[5]

The U-V ratio for manufacturing is 20 percent higher than the ratio for all industries. This is caused by a lower than average unemployment rate and a much lower than average new hire rate. Our research has shown that manufacturing is one of the low turnover industries. It has also shown that construction is a high turnover industry. It tends to have high unemployment rates, but couple that with very high turnover rates and the result is an average U-V ratio. This is an example of an industry with a lot of hiring and a lot of movement in and out of unemployment. Trade and services are examples on a lower scale of labor market turmoil, while finance and transportation join manufacturing as more "stable" industries.

[5]It is unclear that each industry will have a 10-day average duration. The Bureau of Labor Statistics (BLS) study referred to by Abraham does not give duration by industry. (See Lois Plunkert, 1981.) Obviously, the choice of average duration makes a big difference in the U-V ratio. This points out one weakness of constructing a vacancy rate from a new hire rate.

The U-V ratio varies widely by state.[6] In Table 5 it varies from 2.79 in Michigan to less than one in Nebraska. In Nebraska, low unemployment and average turnover combine for the low U-V ratio, while in Michigan it is low turnover and high unemployment that cause the high ratio. Note that the state in this sample with the highest unemployment rate (California) does not have as high a U-V ratio as some of the industrial states since its new hire rate is relatively high. Nevada has the highest new hire rate on the list, but it also has higher than average unemployment and thus does not have the lowest ratio.

There are problems with using vacancy data as an aggregate labor market indicator, especially vacancy data derived from new hire rates. We believe that the new hire rate should be used directly as a measure of labor market activity, rather than vacancy rates. The problems of the vacancy rate surveys are well known and the crucial importance of average duration on the transformation from the new hire rate to a vacancy rate makes the constructed vacancy rate weak.[7] Abraham is able to do little more than speculate on possible average duration at different levels of unemployment.[8] A new hire is a labor market transaction that can be counted; the perception of a vacancy depends on too many subjective factors. In addition, the new hire data can provide follow-up information on duration of the hire. In the past, new hire rates were incomplete. By using the Social Security sample in conjunction with state Unemployment Insurance records to derive new hire rates, we believe that we now have a tool that can be used to enhance labor market research and help to build better labor market policy. Government planners have to have access to more disaggregate data so that they can target programs to the correct area and industry. Different segments of the population are more likely to be affected by structural unemployment than others. It is important that we are aiming at the right people when we design programs to help alleviate unemployment.

[6] Once again, the average duration is assumed to be 10 days for all states. The BLS study by Plunkert showed an average duration of about 10 days for Massachusetts and about 8 days for Utah in 1980. The rates for other states are unknown.

[7] For some of the conceptual problems with the vacancy data see NBER (1966).

[8] In addition, the BLS study showed that in Massachusetts, the average duration varied widely by occupation. Low-skilled jobs stayed open much less time than did jobs for high-skilled, white-collar workers (see Plunkert).

REFERENCES

Abraham, Katherine, "Structural/Frictional vs. Deficient Demand Unemployment," *American Economic Review*, September 1983, *73*, 708–24.

Cohen, Malcolm S. and Schwartz, Arthur R., "U.S. Labor Turnover: Analysis of a New Measure," *Monthly Labor Review*, November 1980, *103*, 9–13.

______ **and** ______, "A New Hires Model for the Private, Non-Farm Economy," in *The Economic Outlook for 1984*, Thirty-First Annual Conference on the Economic Outlook, Ann Arbor, November 1983.

______, ______, **and Grimes, Donald R.,** *The ILIR New Hires Model*, Ann Arbor: ILIR, University of Michigan, 1983.

Plunkert, Lois, *Job Openings Pilot Program: Final Report*, Washington: Bureau of Labor Statistics, 1981.

National Bureau of Economic Research, *The Measurement and Interpretation of Job Vacancies*, Other Conference Series, No. 5, New York: Columbia University Press, 1966.

Structural/Frictional vs. Deficient Demand Unemployment: Reply

By KATHARINE G. ABRAHAM*

The purpose of my 1983 article was to document the relative numbers of unemployed persons and vacant jobs at various unemployment rates. This information should be highly relevant to policymakers: programs designed to improve the process of matching in the labor market cannot greatly lower the unemployment rate if the number of persons desiring work greatly exceeds the number of jobs available, whereas they might be more successful if the unemployed-to-vacancy ratio were lower.

My earlier paper makes use of two distinct approaches to estimating the job vacancy rate. The first procedure corrects data from surveys which asked employers to report their stock of unfilled positions for various sources of downward bias. The second procedure rests on the fact that, in steady state, the vacancy rate equals the new hire rate times the average vacancy duration. New hire rate estimates are constructed using job tenure data from each of four January *CPS*s, then combined with the limited available direct evidence on average vacancy duration to yield vacancy rate estimates.[1] The bulk of the paper is devoted to the set of estimates based on employer surveys; the set based on *CPS* job tenure data is presented as corroborative rather than as compelling in its own right. While either can be criticized, I find it reassuring that two such different approaches to estimating the vacancy rate should yield such consistent results.

Arthur Schwartz, Malcolm Cohen, and Donald Grimes (1986) make no comment on the first of these sets of vacancy estimates, but offer an interesting alternative to the second. In place of the *CPS* data I use to estimate the new hire rate, they use employers' Social Security records; their assumptions about average vacancy duration also differ from mine. Schwartz et al.'s vacancy estimates are somewhat higher than those I report for comparable unemployment rates; their higher vacancy estimates imply somewhat lower unemployment-to-vacancy ratios.

Part of Schwartz et al.'s motivation for substituting Social Security-based new hire estimates for *CPS*-based new hire estimates is their belief that the latter are seriously downward biased. Individuals may hold more than one job over a period, but are counted no more than once in new hire estimates based on the proportion of individuals employed as of the period's end whose jobs started during the period. While Schwartz et al.'s argument on this point is logically correct, it seems unlikely to have much practical significance for the estimates I report. These are based on the proportions of persons employed as of the January *CPS* reference week whose jobs started between January 1 and the end of the reference week. Surely the number of jobs that both start and end during these two or three week intervals is relatively small. In my view, a more serious issue is the fact that only four *CPS* interviews during the 1960's and the 1970's produced job tenure data suitable for estimating the new hire rate.[2] More extensive information on new hires would certainly be very welcome.

*Research Associate, The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036. The views expressed here are my own, and should not be attributed to the trustees, officers, or staff members of The Brookings Institution.

[1] The estimated vacancy rate equals the estimated new hire rate times estimated average vacancy duration.

[2] Job tenure data are also available for May 1979, but the lowest tenure category is "employed one year or less." Schwartz et al.'s argument carries more force when applied to these data: using the proportion of currently employed individuals with tenure in the zero to one-year interval to estimate the new hire rate *would* produce seriously downward biased results.

Unfortunately, there is reason to suspect that Schwartz et al.'s Social Security-based new hire estimates, and thus the associated vacancy estimates, may be upward biased. They identify new hires by checking for new individual Social Security numbers on employers' reports. This means that any time an error is made in entering a continuing employees' Social Security number on an employer's report, or in keypunching a continuing employee's Social Security number at the Social Security office, a new hire is erroneously inferred. There could be a substantial number of this sort of errors. A project carried out by Employment Service personnel in the state of California to determine whether employers' Unemployment Insurance (*UI*) tax contribution reports could be used to construct new hire rate estimates has produced some suggestive statistics. In one study conducted as part of this project, some 300,000 wage records were examined; errors were detected in approximately 27,000 Social Security numbers or 9 percent of the total. In another study involving 6791 wage records supplied by a not necessarily random sample of 55 metropolitan area employers, false changes in employees' Social Security numbers produced 16 percent overreporting of accessions.[3] The *UI* system and the Social Security system are similar in that both rely on employer reporting, both lack a built-in computerized check that detects errors in individuals' Social Security numbers, and both ultimately rely on the claim-filing process to clear errors affecting individuals' benefit eligibility. It thus seems reasonable to suppose that individual Social Security number errors are also a significant problem for data in the Social Security system.

A second concern regarding Schwartz et al.'s new hire estimates is that they rest on a limited time-series of underlying data. While Schwartz et al. present numbers for 1973 through 1982, they have actual Social Security records only for 1973 through 1975. The later years' numbers are projections based on the relationship of the quarterly new hire rate to the percentage change in employment over the quarter, the previous quarter's unemployment rate, and three seasonal dummies, estimated using the available twelve quarters of data.[4]

The second building block for both my *CPS*-based vacancy estimates and Schwartz et al.'s vacancy estimates is an assumption concerning average vacancy duration. I assumed that average vacancy duration ranged between 5 and 15 days; they assume that average vacancy duration always equals 10 days. If average vacancy durations do in fact lengthen in tight labor markets and shorten in loose labor markets, Schwartz et al.'s calculations may well overstate the vacancy rate associated with higher unemployment rates.

Table 1 presents some evidence concerning cyclical movements in average vacancy durations. Average vacancy duration is assumed to equal the vacancy rate divided by the new hire rate.[5] The underlying vacancy rate and new hire rate data came from two employer surveys, the first covering the Minnesota nonagricultural sector over the period January 1972 through December 1979, and the second covering U.S. manufacturing over the period April 1969 through December 1973. These vacancy and new hire estimates are almost certainly biased downward; each vacancy and new hire series was thus multiplied by an appropriate correction factor before creating the estimated average vacancy duration series.[6] Finally, the average duration series was seasonally adjusted.[7]

The estimated average vacancy durations reported in Table 1 do depend upon the

[3]See California Employment Development Department (1979). Malcolm Cohen kindly made this report available to me.

[4]See Cohen and Schwartz (1983).

[5]This follows from the equation specified in fn. 1.

[6]The data and correction factors are described in my earlier paper. The correction factors equal: 2.175, Minnesota vacancy rates; 1.634, Minnesota new hire rates; 2.104, U.S. manufacturing vacancy rates; and 1.581, U.S. manufacturing new hire rates.

[7]The seasonal adjustment procedure consisted of regressing average vacancy duration on 12 month dummies; subtracting the relevant month dummy coefficient from each observation; and then adding back the *average* of the month dummy coefficients to each observation.

TABLE 1—UNEMPLOYMENT RATES AND JOB VACANCY DURATION[a]

Unemployment Rate (*UR*) Range	Mean Vacancy Duration (mos.)	Number of Observations
A. Monthly Data, Minnesota Nonagricultural Sector January 1972–December 1979		
$3.0 < UR \leq 3.5$	.415	2
$3.5 < UR \leq 4.0$	.477	12
$4.0 < UR \leq 4.5$	.426	32
$4.5 < UR \leq 5.0$	.423	18
$5.0 < UR \leq 5.5$	.373	8
$5.5 < UR \leq 6.0$	.350	15
$6.0 < UR \leq 6.5$	.333	8
$6.5 < UR \leq 7.0$	.343	1
B. Monthly Data, U.S. Manufacturing April 1969–December 1973		
$3.5 < UR \leq 4.0$	.450	11
$4.0 < UR \leq 4.5$	.410	1
$4.5 < UR \leq 5.0$	.340	12
$5.0 < UR \leq 5.5$	.301	8
$5.5 < UR \leq 6.0$	.265	20
$6.0 < UR \leq 6.5$	.255	5

[a]The methods and sources used in constructing these estimates are described in the text.

correction factors applied to the underlying vacancy rate and new hire rate data. The average vacancy duration associated with an unemployment rate between 5 and 6 percent equals 10 days in the Minnesota estimates and 8 days in the U.S. manufacturing estimates. The fact that these figures are consistent with the average vacancy durations based on starting and ending states of actual openings discussed in my earlier paper suggests that the correction factors are reasonable. The strong cyclical pattern of the average vacancy durations reported in Table 1 does *not* depend on the correction factors. The Minnesota estimates imply that a 4 percent unemployment rate is associated with an average vacancy duration of about 14 days, while a 6 percent unemployment rate is associated with an average vacancy duration of about 10 days, roughly a 30 percent drop. The U.S. manufacturing estimates imply even larger movements in average vacancy duration, from about 13 days at an unemployment rate of 4 percent to about 8 days at an unemployment rate of 6 percent, roughly a 40 percent drop. While the Table 1 data provide no direct information on average vacancy durations at unemployment rates of 7 percent or greater, the strong negative relationship between average vacancy duration and unemployment over the observed range of unemployment rates suggests that average vacancy durations may fall well below the 10 days Schwartz et al. assume at these higher unemployment rates.[8]

In short, Schwartz et al.'s estimates would be greatly improved if some reliable method for detecting coding errors could be applied to the Social Security data used to estimate the new hire rate; if more current records could be gotten from the Social Security Administration; and/or if better assumptions about vacancy durations were employed. My guess is that Schwartz et al.'s vacancy estimates look larger than mine primarily because the presence of individual Social Security number coding errors leads them to overstate the new hire rate and because they assume that average vacancy duration equals 10 days even at very high unemployment rates.

My intent thus far has been to explain the nontrivial differences between my vacancy estimates and those presented by Schwartz et al. It should be noted, however, that the two sets of numbers do have similar implications for policymakers, so that in a very significant sense Schwartz et al.'s results are quite supportive of my original findings. My work suggests that there are between 5 and 8 unemployed persons per vacant job at an unemployment rate of 7.0 percent and between 2.5 and 4 unemployed persons per vacant job at an unemployment rate of 5.0 percent; Schwartz et al. estimate that the unemployed-to-vacancy ratio equals 3.35 at the 1977 unemployment rate of 7.0 percent and 2.01 at the 1973 unemployment rate of 4.9 percent. While my unemployed-to-vacancy ratio estimates are larger than theirs, both sets of estimates imply that even at unemployment rates of 7.0 percent the number of unemployed persons is several times as large as the number of vacant jobs, so that improvements in the process of labor market matching would have limited potential for

[8]The unemployment rates used in Table 1 were seasonally adjusted using the procedure described in fn. 7.

lowering the unemployment rate. Both also imply that the unemployment rate would have to fall well below 5.0 percent before the number of unemployed persons would equal the number of vacant jobs.

The next part of Schwartz et al.'s comment focuses on two significant limitations of aggregate vacancy data for learning about labor market demand conditions. First, the aggregate vacancy figures reveal nothing about whether the available positions are good jobs or bad jobs. Second, the aggregate vacancy figures may mask significant regional and industrial differences. More detailed job vacancy information would indeed be well worth having. Schwartz et al. use their Social Security data to estimate the proportion of new hires that end up lasting less than two quarters; they also produce by-industry and by-state vacancy estimates. Unfortunately, problems with their aggregate vacancy estimates noted above are relevant here as well. Social Security number coding errors seem particularly likely to inflate the estimated number of short-term hires, and disaggregated vacancy estimates that assume a uniform 10-day average vacancy duration may be quite misleading.

Schwartz et al.'s closing paragraph raises the issues of whether it makes sense to think about measuring job vacancies at all. They argue that too many subjective factors are involved in deciding whether a vacancy actually exists for the concept of a vacancy to be empirically meaningful. This is a view I do not share. Previous efforts to collect vacancy data from employers have typically used some variation on the following vacancy definition: "A current job opening is an existing vacant job that is immediately available for filling and for which your firm is actively trying to find or recruit someone from outside your firm (i.e., a new worker, not a company employee)." The survey forms also typically specify the sorts of efforts which constitute "actively trying to find or recruit": listing orders with public or private employment agencies; notifying labor unions or professional organizations; advertising; running recruitment programs; and interviewing applicants. This all seems fairly straight forward. If certain activities have been undertaken, a vacancy exists; if not, no vacancy exists. There is an obvious parallel between the standard approach taken to determining whether an individual is unemployed and this formulation. Most economists are reasonably comfortable with the unemployment concept; they should be no less comfortable with the vacancy concept.

REFERENCES

Abraham, Katharine, "Structural/Frictional vs. Deficient Demand Unemployment: Some New Evidence," *American Economic Review*, September 1983, *73*, 708–24.

Cohen, Malcolm S. and Schwartz, Arthur R., "A New Hires Model for the Private, Non-Farm Economy," in *The Economic Outlook in 1984*, Thirty-First Annual Conference on the Economic Outlook, Ann Arbor, November 1983, 317–46.

Schwartz, Arthur R., Cohen, Malcolm S. and Grimes, Donald R., "Structural/Frictional vs. Deficient Demand Unemployment: Comment," *American Economic Review*, March 1986, *76*, 268–72.

California Employment Development Department, Systems and Research Section, *Employment Service Potential, Volume II: Technical Development Issues*, report prepared for the U.S. Department of Labor, September 1979.

Labor Supply and Tax Rates: Comment

By CECIL E. BOHANON AND T. NORMAN VAN COTT*

In a recent paper in this *Review* (1983), James Gwartney and Richard Stroup attack what they call the "widespread view" on the marginal tax rate-aggregate labor supply issue that holds that the relationship is ambiguous because of the opposing influences of the income and substitution effects generated by a tax rate change. Gwartney and Stroup contend that this view is flawed because it fails to incorporate the effects that tax rate changes have on government spending. In a setting where the government budget is continuously balanced, higher tax rates in the income-substitution experiment imply that government spending rises by the amount of the income effect.[1] The impact of this additional government spending on labor supply must be included, they argue, if an analysis of the labor supply-tax rate issue is to be complete. Viewing additional government spending as conferring a positive income effect, Gwartney and Stroup write:

> ...[I]f individuals value the expansion in output of public goods emanating from the tax increase more than they value the private goods output that must now be foregone, a positive [overall] income effect will reinforce the substitution effect. Unambiguously, the tax increase will cause an increase in the consumption of leisure and a decline in work effort. [p. 449]

Only if extra government spending is valued less than foregone private income will a net negative income effect emerge that could potentially counter the substitution effect.

Although Gwartney and Stroup are correct in pointing out that a complete analysis of the tax rate-labor supply issue must include the effects of government spending, we shall argue that modeling government spending as an offsetting income effect is misleading in that it is *not* applicable to a major portion of government spending. Specifically, for government spending to be treated as an income effect in a two-dimensional income-substitution experiment, government goods must be *identical* to private goods. This holds, however, only if government spending takes the form of transfer payments which neutrally increases private goods-leisure opportunities.

A simple example should make the point clear. In a two-commodity model of oranges and apples, a tax on oranges affects both orange and apple consumption. Moreover, one can incorporate the effect on orange and apple consumption of using revenue from the income effect implicit in the orange tax to provide free oranges. However, one cannot assert that the level of apple consumption would necessarily be the same had the income effect been used to provide free pears. Replace oranges, apples, and pears with private income, leisure, and government goods, respectively, and Gwartney and Stroup's error becomes obvious.

More formally, consider Figure 1 where income is on the horizontal axis and leisure is on the vertical axis. Suppose an individual faces a pre-tax budget constraint of $0Y$ and chooses a utility-maximizing combination of leisure and income consistent with point A. The imposition of an income tax tilts $0Y$ to $0Y'$ where B is initially chosen. (As drawn, the income effect is dominating the substitution effect.)

Gwartney and Stroup incorporate government spending by shifting $0Y'$ in a parallel fashion by the amount of the income effect in the case where government spending and private spending are valued equally. Using a Slutsky income effect definition for expositional purposes, XX' becomes the relevant budget constraint in the income-substi-

*Department of Economics, Ball State University, Muncie, IN 47306.

[1] For the remainder of this paper we shall for expositional convenience consider the case of an increase in marginal tax rates.

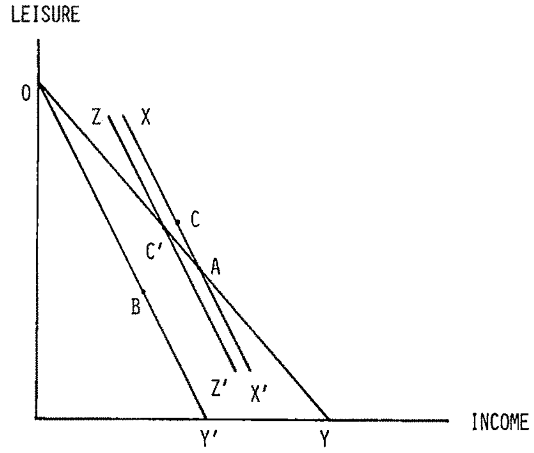


FIGURE 1

tution-spending experiment, and the individual chooses a point such as C. Clearly point C on XX' must be consistent with more leisure consumption than at point A. To the extent tax increases do not, however, confer net additions to capacity the final result will differ quantitatively but not qualitatively from the above conclusion. Point C lies outside the initial opportunity set and therefore cannot be maintained permanently. Final equilibrium occurs when the individual faces an apparent budget constraint such as ZZ' and chooses a point such as point C' which also lies on $0Y$. Only if government spending is valued less than private spending can its impact be less than that of the initial income effect, leaving open the possibility of a decrease in leisure consumption upon the introduction of a tax increase. Gwartney and Stroup argue, therefore, that the ultimate impact of a tax rate change on aggregate labor supply turns on the value of public output compared to private output, not just on the relative strengths of the traditional income and substitution effects.[2]

There is, however, a simple but fundamental flaw in the above analysis: for government spending to shift OY' out in a parallel manner, government spending must be *identical to private income*. Unless one wishes to assert that *all* government spending is a lump sum cash transfer (a rather implausible assertion), modeling government spending as an income effect is simply incorrect. To force other forms of government spending into an income effect analysis would be the analytic equivalent of equating pears and oranges in our previous analogy.

Can the two-dimensional income effect approach be salvaged in the nontransfer case if one uses the income effect as a "proxy" for the value placed on new public spending in terms of foregone private income? No! Consider the apple-orange-pear analogy again. The effect on apple consumption when the orange tax finances pear consumption has nothing to do with the value of pears relative to oranges. Even if the pears provided are valued equally to the potential oranges, final apple consumption will in all likelihood differ between the two expenditure cases because of the different degrees of substitutability (or complementarity) between apples and oranges and apples and pears. Thus, as long as government goods are not identical to private income, the valuation of government goods in terms of foregone private income is not necessarily relevant to the question of aggregate labor supply.[3] For example, if government goods are unrelated to consumption of leisure and private goods in a Hicksian sense, using tax revenue to provide government goods has no effect on the income-leisure choice from point B in Figure 1, even if the spending makes the consumer as well off in the final analysis as at point A with no government goods. The crucial variable in the experiment is not the value of public spending, but the impact of public spending on the labor-leisure choice.

Put another way, the impact of government spending on labor supply from point B in Figure 1 depends upon what those government goods do to the marginal valuation of leisure relative to that of income after their introduction. In the lump sum transfer case,

[2]A mea culpa is in order here. We have essentially made the same argument as Gwartney and Stroup (see our 1984 article). Moreover, the essence of our following critique of the income effect approach is contained in Gordon Winston (1965) and Assar Lindbeck (1982).

[3]We are indebted to Geoffrey Brennan for this point.

government goods are identical to private income. Thus, government provision of private income beyond point *B* quite predictably increases the marginal valuation of leisure relative to that of private income. But as noted previously, this is a special case that has limited applicability.

Finally, one might also note that the criticism of the income effect approach is not limited to the case where government spending finances the provision of purely collective goods.[4] Even if government spending ends up increasing one's real command over both income and leisure, it is not appropriate to characterize such spending as conferring an income effect. If, for example, government spending finances the subsidy of some subset of private goods, the income effect framework is inappropriate because the constancy of the implicit composite private good (income) in Figure 1 is not maintained. Only if government spending preserves the constancy of the composite good of private income *and* maintains the neutrality of the relative valuations of income and leisure is the income effect framework at all applicable.

It is apparent that a complete analysis of a balanced budget tax change on aggregate labor supply must proceed with great care. Although one errs when such an analysis ignores the effect of changes in government spending, it is an equally egregious error to couch those changes as an offsetting income effect, the strength of which depends on the community's valuation of public spending. Instead, the question turns crucially on whether the change in public spending finances the provision of goods that increase or decrease leisure's relative valuation. Unfortunately, this makes the analysis less straightforward. One can speculate as to the impact particular forms of spending have on leisure consumption, but it is quite difficult to assess any particular change in government spending's overall impact.

[4] See Winston who argues that the collective goods nature of government spending precludes the use of an income effect approach.

REFERENCES

Bohanon, Cecil E. and Van Cott, T. Norman, "Shapiro on Marginal Tax Rates and Aggregate Labor Supply: A Comment," *Quarterly Journal of Business and Economics*, Spring 1984, *23*, 15–19.

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–51.

Lindbeck, Assar, "Tax Effects versus Budget Effects on Labor Supply," *Economic Inquiry*, October 1982, *20*, 473–89.

Winston, Gordon, "Taxes, Leisure and Public Goods," *Economica*, February 1965, *32*, 65–69.

Labor Supply and Tax Rates: Comment

By FIROUZ GAHVARI*

In a recent contribution to this *Review*, James Gwartney and Richard Stroup (1983) claim that the "traditional" view on the impact of a change in the wage tax on the supply of labor (namely, the indeterminateness of the direction of the labor supply response) is invalid when it is applied to the economy as a whole. It is the purpose of this comment to demonstrate that Gwartney-Stroup's claim does not stand a closer scrutiny and that the traditional view is in fact correct even when it is applied to the economy as a whole.

I. A Critique of the "Reconstructed View"

Gwartney and Stroup state, quite correctly, that the economy as a whole cannot be made richer by mere tax cuts.[1] From this observation they then go on to argue that:

> For our general equilibrium model, any expansion in expenditures on private goods stemming from the alleged income effect of the tax cut will be exactly offset by a decline in expenditures on public goods (because of the decline in tax revenues which finance them). When the initial level of government expenditure is socially efficient, then at the margin, the additional utility derived from an expansion in consumption of private goods is exactly offset by the decline in utility associated with the reduction in consumption of public goods. The alleged income effect of the tax cut simply disappears. Only the substitution effect remains and *the tax cut will unambiguously increase the quantity supplied of labor.*
>
> [p. 448, emphasis added]

The authors may be correct when they state that the income effect of a tax cut will (for marginal departures from an initially "efficient" level of public expenditure) disappear. They go wrong, however, in their subsequent claim of an *unambiguous* increase in the supply of labor because of the substitution effect. This latter claim would be correct if the tax revenues were always handed back to people as *cash transfers* and if the individuals could buy both private *and* public goods in the market. In that case, a cut in the wage tax would be matched by less compensation. Clearly, then, there would be no income effect associated with such a wage tax cut (except, as noted in fn. 1, for the income effect associated with a reduction in the *existing* excess burden). Moreover, since *all* goods can be bought *freely* in the market, the increase in the price of leisure (as a result of a wage tax cut) relative to *all* other prices will cause the consumption of leisure to go down (supply of labor to rise) and the consumption of other goods to increase. But there is nothing new in this. The tradition has always had it that a compensated wage tax reduction increases the supply of labor.

Gwartney and Stroup, however, do not talk about a compensated wage tax per se. They seem to claim that a tax-cum-public-good-provision policy on the part of the government will be *equivalent*, *for the economy as a whole*, to a compensated wage tax scheme. In their general equilibrium macro model, "Private goods are produced in the market sector and financed by after-tax income. Public goods are supplied by the government and financed from tax receipts" (p. 448). Their point of contention is that, in a general equilibrium setting in which the government uses tax revenues to finance the provision of public goods, there will be no income effect associated with a change in the wage tax, because the change in a worker's take-home pay is exactly matched by the change (of opposite sign) in the supply of public goods to him. In this way, they cancel

*Assistant Professor of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. I thank Barbara Mann, David Salant, and T. N. Tideman for comments.

[1]Apart, of course, from any possible reduction in deadweight losses.

one another and the person's "real" income will remain unchanged. The substitution effect, they then claim, will have to result in an *unambiguous* change in the supply of labor.

The trouble with the above argument is that a change in the *provision* of public goods is *not identical* to a change in the individuals' *purchasing power*. It is true that the income effects of the two changes may be the same; *but* the substitution patterns (because of the tax change), in general, will not. In the latter case, consumers will be *free* to respond to the change in the relative price of leisure to other goods in any which way they wish. They *cannot*, however, do so in the former case. In other words, consumers will be *constrained* in their substitution possibilities. Now since consumers cannot, in this case, attain their desired level of public goods, there can be no a priori presumption regarding the nature and direction of the change in their consumption of other goods *including* leisure.

In order to make this point abundantly clear, assume the government cuts the wage tax (increases the price of leisure). In an unconstrained setting, as will be the case with a concomitant lowering of cash compensations, the substitution effect of this policy will be to *reduce* the consumption of leisure and (most likely) raise the consumption of *all* other goods (private *and* public), assuming public goods are also provided privately.[2] On the other hand, if the wage tax cut is accompanied by a reduction in the supply of public goods, the consumption of public goods *cannot* increase and will have to fall. This constraint, in turn, will cause a different response in the consumption of private goods and leisure (supply of labor) than the predicted one for the unconstrained case.

II. A Simple General Equilibrium Model

Consider a community of n identical individuals each of whom derives utility from consumption of a private good x, leisure l, and a public good G. It is further assumed that the utility function is additive in all goods.[3] Algebraically,

$$(1)\quad u = u(x, l, G) = f(x) + g(l) + h(G).$$

Assume both x and G are produced via linear technologies and there is only one factor of production, namely, labor L. That is,

$$(2)\qquad X = nx = \alpha L_X,$$

$$(3)\qquad G = \beta L_G.$$

These assumptions imply that the relative prices of x, G, and the gross of tax wage will always remain fixed. Moreover, as in Gwartney-Stroup's model, it is assumed that the private good is produced in the market; while the public good is provided by the government and financed out of income (wage) taxes.

Each individual's budget constraint is given by

$$(4)\qquad x = w(1-l)(1-\theta),$$

where each person's endownment of labor is set equal to unity, the price of x is also set equal to unity, w is the gross of tax wage, and θ is the wage tax rate. Moreover, the government's budget constraint is given by

$$(5)\qquad PG = npG = nw(1-l)\theta,$$

[2] To be precise, it is the *expenditure* on all other goods that will rise. In order to ensure that the consumption of *both* private and public goods will increase, it is necessary to employ other assumptions. One sufficient condition is that both public and private goods are normal and gross substitutes to leisure. Another sufficient condition is the normality of all goods plus additivity of the underlying utility function. It must also be noted that in a general equilibrium setting all prices become endogenous. The above conditions are then also based on the assumption of fixed producer prices (linear technologies).

[3] It must be emphasized that the essential point of this comment, namely, the ambiguity of the labor supply response to a change in the wage tax, does not depend on this additivity assumption. It is only employed to simplify the derivations.

where P is the "producer" price of one unit of the public good and $p \equiv P/n$ is its price to each individual.

It is now clear from equation (5) that for each value of θ, a certain amount of G will be provided by the government. This quantity, in turn, will be treated as *given* by consumers in this economy. Each individual maximizes the utility function (1) subject to his budget constraint (4). This yields[4]

$$u_l = w(1-\theta)u_x, \tag{6}$$

where u_i $(i = x, l)$ indicates the first partial derivative of u with respect to good i. Equations (4)–(6) will then determine the amounts of x, l, and G as functions of θ.

In order to analyze the effects of a change in θ on the quantities of x, l, and G, the system of equations (4)–(6) is totally differentiated with respect to θ. As a part of the solution to this new system of equations, the total derivative of l with respect to θ can be found to be equal to

$$\frac{dl}{d\theta} = \frac{wu_x + w^2(1-l)(1-\theta)u_{xx}}{-\left[w^2(1-\theta)^2 u_{xx} + u_{ll}\right]}, \tag{7}$$

where u_{ii} refers to the second partial derivative of u with respect to good i.

It is now assumed that x, l, and G are all normal goods. This assumption, along with the common axioms on the preference ordering, will imply diminishing marginal utility for all goods, namely, $u_{ii} < 0$. The sign of the denominator in (7) is then unambiguously positive. The sign of the numerator, on the other hand, is ambiguous and consequently $dl/d\theta$ cannot be signed.[5]

It is emphasized that this ambiguity in sign in general exists whether or not the initial supply of the public good is second-best efficient,[6] namely, when θ is initially set by the government such that $du/d\theta = 0$. Indeed, for the special case of a logarithmic utility function it can easily be checked that $dl/d\theta$ is always (including at the second-best efficient value of θ) equal to zero.[7] That is, the constrained substitution effect and the (possible) income effect of a tax-cum-public-good-provision policy will just offset one another as is the case in the traditional analysis. This is in contrast with a compensated wage tax scheme that results in $dl/d\theta > 0$.

This, then, reestablishes the traditional view, namely, the indeterminateness of the direction of the labor supply response to a change in the wage tax, in the setting of a *general equilibrium* macro model, in which the tax revenues are not thrown away but are used to finance the provision of public goods.

[4]Alternatively, if it is assumed that each individual "realizes" that the supply of public good by the government is given by equation (5) and *incorporates* this in his optimization problem, the first-order condition (6) will have to be replaced by

$$u_l = (w\theta/p)u_G + w(1-\theta)u_x. \tag{a}$$

[5]If we work with equation (a) instead of (6), the solution would be equal to

$$dl/d\theta = \left[(w/p)u_G - wu_x - w^2(1-l)(1-\theta)u_{xx} + (w/p)^2(1-l)\theta u_{GG}\right] \Big/ \left[w^2(1-\theta)^2 u_{xx} + u_{ll} + (w\theta/p)^2 u_{GG}\right].$$

Again, the sign of $dl/d\theta$ cannot be determined. Note that in this case, the government can choose θ such that the supply of the public good will in fact be first-best efficient. In such a setting the wage tax would not be distortionary. For example, in the special case of a logarithmic utility function ($u = \ln(x) + \ln(l) + \ln(G)$), the equilibrium values of the goods are found to be $l = 1/3$, $x = 2w(1-\theta)/3$, and $G = 2w\theta/3p$. Substituting these values in the utility function and maximizing with respect to θ results in $\theta = 1/2$ so that $l = 1/3$, $x = w/3$, and $G = w/3p$, which are the first-best efficient values for the goods. (The first-best efficient quantities are found by maximizing the above utility function subject to the economy's production possibility constraint, $X + PG = wn(1-l)$.)

[6]Clearly, with distortionary taxes, the supply of public goods cannot be first-best efficient. See also fn. 4 above.

[7]It would be instructive to consider the special case of a logarithmic utility function in more detail. The equilibrium values of the goods, in this case, are found to be $l = 1/2$, $x = w(1-\theta)/2$, and $G = w\theta/2p$. Whether or not the government sets θ at its second-best efficient value (which can be checked to be $\theta = 1/2$ by substituting the above values of l, x, and G in the utility function and then maximizing with respect to θ), it is clear from these solutions that when θ is cut and leisure is made more expensive, instead of the "expected" decrease in l and concomitant increases in *both* x and G, in this *constrained* case where *G has to fall*, only x will go up while l will remain unchanged.

III. Concluding Remarks

This comment indicates that, even for the economy as a whole, the response of labor supply to a wage tax change cannot unambiguously be determined. This was demonstrated for a simple general equilibrium macro model of the economy, where the government uses the revenues from its taxation of wage income to provide a public good. It was also argued that Gwartney and Stroup's misconception arises out of their equating the effects of a change in the government's provision of public goods with those of outright cash subsidies in an economy where public goods are provided privately. This identification, however, was shown to be an unwarranted one, because of the constrained nature of substitution possibilities in the former case.

REFERENCE

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–51.

Labor Supply and Tax Rates: Reply

By JAMES GWARTNEY AND RICHARD L. STROUP*

The comments of Firouz Gahvari and of Cecil Bohanon and T. Norman Van Cott provide useful extensions of our earlier analysis. Nonetheless, our central point remains intact: the traditional labor-leisure analysis is invalid because it ignores the effects of changes in government spending on individual welfare.

Gahvari points out that, in the case of public goods, the linkage between changes in tax rates and labor supply is more complex than we implied. Individuals, unable in the large number case to transform leisure into public goods, will be affected differently when the government provides a public good rather than an income transfer. In Gahvari's world, where government goods are irrelevant to all private decisions (complete separability in the utility functions and no ability to purchase public goods privately), the quantity of government goods can be safely ignored in the analysis of private decisions. The decision proceeds as it would if tax revenues were totally wasted, even though the citizens' *total utility* is assumed constant when tax revenues change, with changes in government goods exactly offsetting the utility impacts of the change in private goods. However, we think it is misleading to label the ambiguously signed element beyond the substitution effect as an income effect. How can there be an "income effect" when total utility remains constant?

As the quotes cited by Gahvari from our initial paper (p. 447) illustrate, this is not the income effect envisioned by the traditional work-leisure analysis, which refers to a change in "standard of living" (utility) and reflects the notion that a tax cut will encourage individuals to work less (consume more leisure) by giving them a higher standard of living through more after-tax pay.[1] Implicitly, this view ignores the negative impact on living standards associated with the reduction in the supply of government provided goods (or income transfers).

Gahvari recognizes that when government goods replace private goods such as public education, medical services, food stamps, or cash (and ignoring any cross elasticities), our original analysis stands and there is only the substitution effect.

Bohanon and Van Cott make another refinement, pointing out some secondary effects of government's tax-transfer activity. While changes in tax rates and in government-provided goods will influence the individual's budget constraint, or *ability* to trade off among goods, a complete analysis must also account for the fact that any such shift will move the individual into a new region of his indifference surface. His *willingness* to trade off among the goods may well change. The individual's view of the substitutability or complementarity among government-provided goods, private goods, and leisure becomes relevant. If the government good is strongly enough complementary to leisure and/or substitutable for private goods, the standard substitution effect could indeed be overcome.

Returning to our original paper, we reiterate its central point. The "income effect" component of traditional work-leisure analysis for an individual ignores the individual's utility derived (foregone) from increased (decreased) government spending accompanying changes in revenues. To treat that indi-

*Florida State University, Tallahassee, FL 32306, and Montana State University, Bozeman, MT 59715, respectively. This work was supported by the Political Economy Research Center, Bozeman.

[1]Note the internal contradiction here. If people work and produce *less* because of the alleged income effect accompanying the tax cut, *ceteris paribus*, they will be poorer, not richer.

vidual as typical, and to say that when tax rates are cut, "An increase in real take-home pay is an inducement to work more hours per year, but since it makes workers better off anyway, it may encourage them to take more time off,"[2] is to imply that government expenditures have less value to individuals than private expenditures of the same size.

While models implying a positive relationship between tax rates and quantity of work can be developed, the source of the offsetting factors to the work-leisure substitution effect is not the income effect implied by the traditional two-dimensional work-leisure analysis. The traditional analysis is not vindicated.

[2]See James Tobin (1982, p. 134).

REFERENCES

Bohanan, Cecil E. and Van Cott, T. Norman, "Labor Supply and Tax Rates: Comment," *American Economic Review*, March 1986, *76*, 277–79.

Gahvari, Firouz, "Labor Supply and Tax Rates: Comment," *American Economic Review*, March 1986, *76*, 280–83.

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–51.

Tobin, James, "Supply-Side Economics: What Is It? Will It Work?," *Economic Outlook USA*, Survey Research Center, University of Michigan, Summer 1981; reprinted in *Viewpoints on Supply-Side Economics*, Richmond: Robert F. Dame, Inc., 1982, 132–38.

Social Security and Household Savings: Comment

By George Briden and John Zedella*

In their study in this *Review* (1983), Errki Koskela and Matti Virén reach the conclusion that "... *Social Security variables ... have no effect on the household saving ratio ...*" (p. 215) in an international pool of cross-sectional and time-series data. The purpose of this comment is to demonstrate that their results are not robust. In particular, we shall show that the results are sensitive to sample selection and estimation procedure, and, moreover, that there is indeed evidence in the data which could be interpreted as supporting the "Social Security depresses savings" hypothesis.

I. The Koskela-Virén Approach

In the model that Koskela and Virén estimate, the savings rate is determined within a "gradual adjustment" context. In addition to the lagged savings rate and other regressors, three "Social Security" variables are included in the model. These are SS = the ratio of Social Security benefits to income; OLD = the proportion of people over age 65 in the population; and PR = the labor force participation rate of the older generation. The variable SS is a proxy for the working generation's expected benefit payments, while PR is a proxy for the "induced retirement" effect of Social Security.

The models are estimated by Koskela and Virén using the least squares with dummy variables technique ($LSDV$). The interesting result is the failure of any of the three Social Security variables to be significant at any reasonable level. Thus, the results fail to support the Social Security depresses savings hypothesis.

*Department of Finance and Insurance, College of Business Administration, University of Rhode Island, Kingston, RI 02881. We acknowledge the assistance of Errki Koskela, who generously provided us with access to his data. Blair Lord and Ghon Rhee provided helpful comments.

II. A Reexamination of the Evidence

It is our contention that three factors have combined to reduce the significance of the Social Security variables in the Koskela-Virén study.

1) *Multicollinearity*. There are good reasons, both intuitive and empirical, to expect a high degree of intercorrelation among the Social Security variables used by Koskela-Virén. One intuitive explanation revolves around the political power of those eligible for benefits. The higher the proportion of the population over age 65 (OLD), the more likely that they would be able to use leverage at the polls to pry higher benefits out of elected governments. Alternatively, one might expect that the participation rate (PR) of the population eligible for benefits would depend quite strongly on the rate of benefit payments. The presence of multicollinearity is suggested by Koskela and Virén when they report an R^2 of .99 when the variable PR is regressed on OLD, SS, and other variables.[1] The t-tests upon which they rely to support their conclusions are therefore of questionable power.

2) *Measurement error*. The variable PR has many missing values, a problem which Koskela and Virén solve via extrapolation. The coefficient of PR is now subject to a bias of unknown direction, and the power of Koskela and Virén's hypothesis tests can once again be expected to be impacted adversely.

3) *Efficiency*. The inefficiency of the $LSDV$ method is well known. Degrees of freedom are lost when dummy variables are included, and the method ignores "between group" variation. A "lack of significance" argument such as Koskela and Virén's is thus less convincing when it is based upon the results of an $LSDV$ estimation.

[1] See their Table 1, p. 214. Note the t-statistics for SS and OLD in col. 3.

TABLE 1—COMPARISON OF REGRESSION RESULTS PERSONAL SAVINGS RATE MODELS

	(1)	(2)	(3)	(4)	(5)
$q(t)$	.1639	.1929	.1768	.0391	.0409
	(5.26)	(5.14)	(5.15)	(2.92)	(3.28)
$g(t)$	.3816	.4153	.3771	.3235	.324
	(11.94)	(12.53)	(11.09)	(24.55)	(25.22)
$s/y(t-1)$	.5771	.5707	.5983)	.9563	.9554
	(13.28)	(11.61)	(11.82)	(117.52)	(131.42)
$r(t)$	.1184	.1427	.0489	−.006	.0047
	(1.48)	(1.70)	(.63)	(.23)	(.21)
$\Delta U(t)$	.0043	.0046	.0054	.0036	.0033
	(3.51)	(3.87)	(4.53)	(8.17)	(8.23)
$SS(t)$	−.004	−.001	−.005	−.005	−.005
	(.74)	(.21)	(.81)	(4.88)	(5.92)
$OLD(t)$	−.222	−.524	−.701	.0365	
	(1.16)	(2.66)	(3.41)	(2.51)	
$PR(t)$	−.032	−.070	−.123	.0057	
	(.52)	(1.26)	(1.68)	(.49)	
N	240	162	182	182	182
R^2	.96	.97	.97	–	–

Notes: Absolute values of *t*-statistics are shown in parentheses. Cols. 1 through 3 were estimated using *LSDV*. The individual intercepts are deleted for the sake of brevity. All variables are defined as in Table 1 of Koskela and Virén. Col. 1 uses their data set in its entirety. Col. 2 is based on a sample for all countries with complete data for the period 1960–77. Col. 3 is based on a sample of all countries with complete data from 1964–77. Cols. 4 and 5 are based on the same sample as col. 3, but are estimated using the Parks method.

Given the above set of circumstances, it would not be surprising to discover that the results are sensitive to sample, specification, and estimation procedure. In fact, this is the case.

Table 1 presents typical results for a variety of samples. Column 1 is based upon the same data used by Koskela and Virén and is provided for comparison. Columns 2 and 3 of are based upon subsamples which are symmetric, which is to say that each country included in the pool is restricted to the same time period.[2] The selection of countries was quite arbitrary, with the only criterion being the desire to obtain the largest possible symmetric subsamples.

Note that as the sample is varied, considerable variability both in coefficient magnitude and level of significance is apparent for the suspect variables *SS*, *OLD*, and *PR*. This is exactly what would be expected if the above described problems were contaminating the data.

We also experimented with more efficient estimation procedures. The procedure selected for this exercise is due to Richard Parks (1967), and the results are presented in columns 4 and 5 of Table 1. The Parks estimator is a generalized least squares procedure which allows for the cross-correlated, heteroscedastic residuals which might occur in this case.[3] The use of this more efficient estimation procedure produces a significant coefficient for *SS* regardless of whether or not the variables *OLD* and *PR* are included. Similar results are obtained using other samples and estimation techniques.

[2] This sort of sample is convenient if other estimation techniques available for "pooled" samples (for example, the error components method) are to be used.

[3] In fact, when the sample is divided into two groups on the basis of average per capita Social Security benefits, the likelihood ratio test for homoscedasticity yields a calculated *Chi*-square of 450.6. Heteroscedasticity is thus indicated. It is also apparent from this exercise that pooling the data is inappropriate in this case. The coefficient estimates are widely divergent across samples, suggesting another possible reason for the Koskela-Virén results.

The results demonstrate that Social Security's estimated impact on savings is sensitive to the estimation procedure employed, the sample, and the model specification. Moreover, depending upon which estimator is used, it might be said that there is, in fact, evidence on this data which supports the Social Security depresses savings hypothesis.

REFERENCES

Barro, Robert J. and MacDonald, Glenn M., "Social Security and Consumer Spending in an International Cross Section," *Journal of Public Economics*, June 1979, *11*, 275–89.

Koskela, Errki and Virén, Matti, "Social Security and Household Saving in an International Cross Section," *American Economic Review*, March 1983, *73*, 212–17.

Parks, Richard W., "Efficient Estimation of a System of Regression Equations when Disturbances are Both Serially and Contemporally Correlated," *Journal of the American Statistical Association*, June 1967, *62*, 500–09.

Social Security and Household Saving: Reply

By Erkki Koskela and Matti Virén*

George Briden and John Zedella suggest that the results in our 1983 paper about the effect of Social Security on household savings are not robust. More specifically, they think that the significance of Social Security variables in our study has been reduced by 1) multicollinearity, 2) measurement errors, and 3) inefficiency of the *LSDV* method. They go so far as to claim that "there is, in fact, evidence on this data which supports the Social Security depresses savings hypothesis" (p. 288). We do not concur with their assessment and see no reason to change our earlier conclusions.

1) Briden and Zedella claim that the Social Security variables *SS*, *OLD*, and *PR* are highly multicollinear thus making t-tests of questionable value. Their example in this context is misleading, to say the least; the high "explanatory power" of the participation equation is mainly due to the lagged dependent variable and country dummies, not to the *SS* and *OLD* variables (see our earlier paper, p. 214). Nevertheless, we checked whether the results with the pooled cross-country data were affected by using various ridge estimation techniques, which allow for potential multicollinearity. Without exception, no qualitative changes could be detected; in particular, the Social Security variables were insignificant in all cases (see our forthcoming 1986 paper).

2) Undoubtedly, measurement errors may be a problem here (and in most other contexts as well). But the statement by Briden and Zedella is again a bit misleading. Under the usual assumptions about the nature of measurement errors, the potential measurement error of the *PR* variable will give the downward biased *OLS* estimate to it. On the other hand, the effect of this potential measurement error on the parameter estimate of the *SS* variable is not a priori unambiguous, even when all other variables are assumed to be measured without error. On the other hand, if more than one of the independent variables are measured with error, nothing can be said about the nature of the bias unless we have information about the variance-covariance matrix of measurement errors—a most unlikely event (see, for example, Henri Theil, 1971).

3) In order to evaluate the robustness of our results, Briden and Zedella present estimation results on a variety of samples and experiment with the Parks estimation procedure, which allows for both contemporaneous and first-order serial correlation of disturbances. They wind up with the verdict that our results are robust with respect to neither the sample selection nor the *LSDV* estimation method.

Let us take the sample selection first. Even though there seems to be some variation in the coefficient estimates as the sample is varied, it should be pointed out that the null hypothesis that the coefficient of the *SS* variable is zero—the major variable of interest in our earlier study—cannot be rejected at any standard levels of significance in subsamples used by Briden and Zedella (see the first three columns of their Table 1).[1]

Let us turn to the estimation results obtained by Briden and Zedella using the Parks procedure. These estimation results, pre-

*Department of Economics, University of Helsinki, and Bank of Finland, P.O. Box 160, 00101 Helsinki, Finland, respectively. We acknowledge financial support from the Yrjö Jahnsson Foundation.

[1]The estimation results with the individual cross-country data suggested roughly the same; the null hypothesis that the coefficient of the *SS* variable is zero in the savings rate equation could be rejected only in the cases of Portugal and Sweden at the 5 percent but not at the 1 percent significance level, irrespective of whether the standard t-ratios or the adjusted t-ratios (corrected by the Halbert White 1980 procedure to account for potential heteroscedasticity) were used (see our 1986 paper).

sented in columns 4 and 5 of their Table 1, constitute the only evidence that seems to challenge the robustness of our conclusion that the Social Security depresses saving hypothesis receives no support. In this case, the Parks procedure produces strikingly precise coefficient estimates, so that, for example, even though the coefficient estimate of the *SS* variable remains practically the same in the savings rate equation, its *t*-ratio increases from .81 to 4.88 (see column 4 of their table).[2] This implied increase in efficiency does surprise us very much. However, for a number of reasons, we do not regard this—even though a little puzzling—as constituting serious evidence against the robustness of our results in our earlier paper.

Briden and Zedella's explanation that it is precisely heteroscedasticity that substantially biases our estimates of standard errors is not convincing. When the *t*-ratios were adjusted by the White procedure to account for (potential) heteroscedasticity, the resulting changes in the *t*-ratios were not worth mentioning. In particular, the adjusted *t*-ratios did not make the coefficient estimate of the *SS* variable significantly different from zero (see our earlier paper for details). Also, for some reason, the Parks procedure produces the very high value of the coefficient estimate of the lagged savings ratio, namely .956 in contrast with .577 obtained with the *LSDV* method. Comparing this estimate with all previous studies and with the individual cross-country results from the same sample shows that this coefficient estimate is completely out of line with these.[3] It implies both the unreasonably long adjustment period (according to estimates it takes approximately 23 years for the savings ratio to adjust to the long-run desired level!) and the unreasonably high long-run effects of explanatory variables on the savings ratio; for instance, the long-run effect of a 1 percent increase in real income on the savings ratio is over 7 percent! Since we cannot regard these properties of the savings ratio equation as plausible in any sense, neither can we believe that this evidence supports the "Social Security depresses savings" hypothesis.

[2]In fact, the standard errors of the coefficients decrease to such extent that—except for r_t and ΔU_t—the respective standard errors are all .01!

[3]With individual country data, the mean and standard deviation of the coefficient estimates of $(s/y)_{t-1}$ are .411 and .342, respectively.

REFERENCES

Briden, George and Zedella, John, "Social Security and Household Savings: Comment," *American Economic Review*, March 1986, *76*, 286–88.

Koskela, Erkki and Virén, Matti, "Social Security and Household Saving in an International Cross Section," *American Economic Review*, March 1983, *73*, 212–17.

______ **and** ______, "Social Security and Household Saving in an International Cross Section: Some Further Evidence," *Finnish Economic Journal*, 1986 forthcoming.

Theil, Henri, *Principles of Econometrics*, Amsterdam: North-Holland, 1971.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

Notes

The eighth World Congress of the International Economic Association (IEA) will be held in New Delhi, India, December 1–5, 1986. The theme is Balance Between Industry and Agriculture in Economic Development. Those interested in presenting a paper should send proposals with abstracts to Professor Kenneth J. Arrow, IEA President, Department of Economics, Stanford University, Stanford, CA 94305. Requests for information should address IEA, 23 Rue Campagne Premiere, 75014 Paris, France.

Corruption and Reform is a new journal publishing comparative research in the areas of political corruption, political finance, and policy analysis of reforms, three times a year, beginning 1986. Each issue will contain research articles and book reviews, news of research meetings and conferences, and short reports on important events, legislation, and investigations in various countries. Papers employing the methods and perspectives of economics are invited, as are manuscript referees, and book reviewers. Manuscripts (4 copies) and inquiries may be directed to Co-Editors Michael Johnston, Department of Political Science, University of Pittsburgh, Pittsburgh, PA 15260 (telephone 412 + 624–3702), or Stephen P. Riley, Department of International Relations and Politics, North Staffordshire Polytechnic, College Road, Stoke-on-Trent ST4 2DE, England (telephone 0782–45531, ext. 343). Books for review and offers to serve as reviewer should be sent to Alan Doig, Faculty of Social and Environmental Studies, Roxby Bldg, University of Liverpool L69 3BX, England (telephone 051–709–6022, ext. 2755).

The sixth International Conference on Decision Support Systems (DSS-86) will be held in Washington, D.C., April 21–24, 1986. It will provide a forum for learning about and exchanging experiences and ideas on Decision Support Systems. Contact Julie Eldridge, DSS–86, 290 Westminster Street, Providence, RI 02903.

Call for Contributors: The Handbook of American Business History is to be a compilation of histories and bibliographies of American business sectors, to be organized along the lines of the Enterprise Standard Industrial Classification. Contact David O. Whitten, Editor, *The Handbook of American Business History*, Department of Economics, Auburn University, Auburn, AL 36849–3501.

Call for Papers: The 1986 meetings of the Pennsylvania Economic Association will be held May 30–31, at Gettysburg College. Abstracts for papers should be sent before March 15 to William N. Ross, Vice-President, PEA Program, Department of Economics, Clarion University, Clarion, PA 16214. For further information, contact Derrick K. Gondwe, Department of Economics, Gettysburg College, Gettysburg, PA 17325.

Call for Papers: Análisis Ecónomico, a new bilingual journal sponsored by the University of Santiago, will be published in June and November each year. Manuscripts in all fields are sought, but preference will be given to theoretical, empirical, and policy-oriented papers on *LDC*s. Three copies of the paper and a 100-page abstract, in English or Spanish, should be submitted to the Editor, *Análisis Ecónomico*, Departmento de Economia, Universidad de Santiago, P.O. Box 4637, Santiago 2, Chile.

Call for Papers: The annual meeting of the European Finance Association will be held August 28–30, 1986, in Dublin. Send proposed papers or abstracts by April 1 to Michael Walsh, Department of Banking and Finance, University College Dublin, Belfield, Dublin 4, Ireland (telephone 69 32 44, ext. 8214). For general information, contact Ms. Gerry Dirickx-Van Dyck, EFMD–EIASM, 13 Rue d'Egmont, B–1050 Brussels, Belgium.

Call for Papers: The fourth annual meeting of the Association of Managerial Economists will be held in New Orleans, LA, December 28–30, 1986, in conjunction with the ASSA meetings. Three sessions of contributed papers will be featured. Areas include advertising, competitive strategy, diversification, financial decisions, forecasting, innovation, managerial labor markets, market for corporate control, and pricing, etc. Both members and nonmembers are invited to send papers and/or program suggestions to Professor Mark Hirschey, AME Program Chairman, Graduate School of Business Administration, University of Colorado, Denver, CO 80200 (telephone 303 + 623–4436).

Call for Papers: The AEA Committee on the Status of Women (CSWEP) will sponsor two sessions at the Southern Economic Association meetings, November 23–25, 1986, in New Orleans: "Pay Equity: Theory and Applications" (send abstracts to Marie Lobue, Economics and Finance Department, University of New Orleans,

LA 70148); and "Occupational Segregation: Issues and Analyses" (send abstracts to Lavonia Casperson, Department of Economics and Finance, Louisana State University, 8515 Youree Way, Shreveport, LA 71115).

Call for Papers: The thirteenth annual conference of the European Association for Research in Industrial Economics (EARIE) will be held August 24–26, 1986, in West Berlin. Sessions include the economics of industry and firm organization, industrial and competition policy, industrial organization and international trade, regulation of industries and firms, industrial organization theory, technological and organizational change and industry structure, as well as other areas of current research interest. To present a paper, send a one-page abstract and three copies of the paper no later than April 1 to J. Müller, Programme Chairman, EARIE Conference, c/o German Institute of Economic Research, Koenigin-Luise-Str.5, 1000 Berlin 33, West Germany (telephone 49–30–82991 328 or 829910).

Call for Papers: Social Science Microcomputer Review was begun in 1982 as a quarterly forum on the research and teaching applications of microcomputing. In 1985, its scope was expanded to include economics. Articles, tutorials, book and software reviews, news, and comments of interest to economists or other social scientists are invited. Style requirments and other information may be obtained from the editor-in-chief, G. David Garson, Box 8101, North Carolina State University, Raleigh, NC 27695, or from the associate editor for economics, William P. Yohe, Department of Economics, Duke University, Durham, NC 27706.

Call for Papers: The Department of Economics and Management Sciences at Eastern Connecticut State University announces the establishment of the David T. Chase Free Enterprise Institute. Among other activities, the Institute will issue a biannual journal that publicizes the research results and reflections of academics and business professionals. Manuscripts are sought that address relevent and topical issues of a domestic or international nature in economics, or the various functional areas of business. Authors should also send a brief biography and short abstract to Dr. Kenneth M. Parzych, Editor, Chase Free Enterprise Institute, Eastern Connecticut State University, Willimantic, CT 06226–2295.

The third National Conference on Environmental Dispute Resolution will be held in Washington, D.C., May 29–30, 1986. Sponsored by The Conservation Foundation, it is a conference for business leaders, environmentalists, public officials, planners, and others interested in new approaches for resolving environmental disputes. For further information, contact Gail Bingham, Senior Associate, The Conservation Foundation, 1255 23rd Street, NW, Washington, D.C. 20037 (telephone 202+293–4800).

The eighth annual Quebec Summer Seminar, organized by the Center for the Study of Canada at SUNY-Plattsburgh, will be held in Montreal and Quebec City, May 31–June 7, 1986. The program will involve extensive interview/discussions with Quebec academics, politicians, media personalities, business people, and cultural elites; it also includes a concert by the Montreal Symphony Orchestra, a reception hosted by the Quebec government, a luncheon at a rural Quebec Inn, and tours of Montreal and Quebec City. The program will assume all costs for lodging, some meals, in-province transportation, cultural events, and tours. A registration fee of $175.00 will be charged. All full-time academics at U.S. universities who can demonstrate that the information acquired at the seminar will be incorporated in their courses and/or will assist them in research projects, will be eligible to apply. Approximately 30 people will be selected. The application deadline is March 8, 1986. For more detailed information and application forms, please contact 8th Annual Quebec Summer Seminar, Center for the Study of Canada, SUNY-Plattsburgh, Plattsburgh, NY 12901 (telephone 518+564–2086).

Public Domain Software: A floppy disk of computer programs designed primarily for instructional purposes that can be freely copied for use by colleagues and students has been developed at Wesleyan University. Included is a program for simulating the short- and long-run consequences of alternative monetary and fiscal policies and a program for solving linear programming problems (simplex algorithm). A disk for use on the Dec Rainbow, Osborne, or other CP/M machines may be obtained for a nominal charge from Mike Lovell, Department of Economics, Wesleyan University, Middletown, CT 06457.

The U.S. Department of Labor has contracted with the Social Science Research Council to appoint an advisory group to plan a possible 1986 Quality of Employment Survey. Researchers with an interest in working conditions, job satisfiaction, productivity, labor management, and other data that might be produced if such a survey were mounted are invited to communicate their interest and ideas to Richard C. Rickwell, Social Science Research Council, 605 Third Avenue, New York, NY 10158.

The National Institute of Mental Health is now accepting applications for research grants on reimbursement issues in mental health services delivery. This program is part of NIMH's Mental Health Economics Research Program, designed to increase research studies specifically analyzing a variety of reimbursement policies and issues that reflect the NIMH's recognition of

the importance of reimbursement policies and issues in developing and expanding mental health services to all persons in need. Studies are sought that have implications beyond the immediate research setting/data base. Priority will be given to applications that demonstrate sophistication in economic analysis and sensitivity to institutional issues and unique characteristics of mental health sectory. Copies of this grant announcement (MH-86-08) and information can be obtained from Mr. Paul Widem, Chief, Mental Health Economics Research Program, Biometric and Applied Sciences, NIMH, Room 18 C-03, 5600 Fishers Lane, Rockville, MD 20857 (telephone 301+443-4233).

The Inter-University Consortium for Political and Social Research (ICPSR) will hold its 1986 Summer Program in Quantitative Methods of Social Research in Ann Arbor, June 30-August 22. The Program is divided into two 4-week sessions: June 30-July 25, and July 28-August 22. Individuals can attend either or both, or participate in one or more of the shorter workshops. For course offerings and full information, contact ICPSR official representatives at member colleges and universities, or Henry Heitowit, Director, Educational Resources, ICPSR Summer Program, P.O. Box 1248, Ann Arbor, MI 48106 (telephone 313+764-8392).

The twenty-first International Atlantic Economic Conference, "Charting New Frontiers," will be held in St. Thomas, Virgin Islands, April 15-20, 1986. For full information, contact John M. Virgo, International Program Chairman, Atlantic Economic Conference, Southern Illinois University, Edwardsville, IL 62026-1101.

The International Association for Community Development (IACD) will hold an international colloquium, "Mobilisation of Human Resources and Community Development," October 13-18, 1986, in Marcinelle. For further information, contact IACD, 179 rue du Debarcadere, 6001 Marcinelle, Belgium (telephone 0+71/43 20 72; 43 31 83; 36 62 73).

The Council for International Exchange of Scholars (CIES) announces availability of 1986-87 Fulbright Lecturing Grants to U.S. faculty in the field of economics. Specific openings are in Argentina, Botswana, Bulgaria, Burundi, China, Colombia, Egypt, Fiji, Gabon, Hungary, Jamaica, Japan, Peru, Malawi, Mexico, Niger, Nigeria, Pakistan, Papua New Guinea, Jordan, Philippines, Poland, Senegal, Somalia, Sudan, Swaziland, Tanzania, Thailand, Tunisia, Turkey, USSR, and Zambia. Faculty in all academic ranks, including emeritus, and independent scholars are eligible. Applicants should have a Ph.D., college or university teaching experience, and reasonable evidence of scholarly productivity; U.S. citizenship is required. For information contact CIES, Eleven Dupont Circle, NW, Suite 300, Washington, D.C. 20036 (telephone 202+939-5401).

The Association for the Advancement of Policy, Research and Development in the Third World announces two conferences on Science, Technology, and Industrialization in the Third World. The first will be held June 5-7 in Brussels, Belgium. The proposal deadline is April 15. Contact Dr. Shah M. Mehrabi, Program Coordinator, Department of Economics, Mary Washington College, 1301 College Avenue, Fredericksburg, VA 22401 (telephone 703+899-4092). The second will be held September 25-28 in Berkeley, California. The proposal deadline is July 20. Contact Dr. Mekki Mtewa, Executive Director, AAPRD, 201 rue Belliard, Box 14, B-1040 Brussels, Belgium.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45 Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Death

John H. Keith, Jr., senior economist, Data Resources, Inc., September 18, 1985.

Retirements

O. Ralph Blackmore: professor emeritus of economics, Wilfrid Laurier University, July 1985.

Robert Campbell: professor of economics, University of Oregon, December 31, 1985.

H. T. Koplin: professor of economics, University of Oregon, June 15, 1985.

Stefan H. Robock, Robert D. Calkins professor of international business, Columbia University, July 1985.

Promotions

M. Akbar Akhtar: vice president and assistant director of research, Federal Reserve Bank of New York, July 19, 1985.

Mark C. Berger: associate professor, department of economics, University of Kentucky, July 1, 1985.

Richard J. Butler: associate professor of economics, Brigham Young University, September, 1985.

Peter J. Eaton: associate professor of economics, University of Missouri-Kansas City.

A. Steven Englander: research officer and senior economist, Research and Statistics Function, Federal Reserve Bank of New York, July 19, 1985.

Edward J. Frydl: vice president and assistant director of research, Federal Reserve Bank of New York, July 19, 1985.

Kenneth J. Guentner: manager, securities department, Federal Reserve Bank of New York, July 19, 1985.

L. Jay Helms: associate professor of economics, University of California-Davis, July 1, 1985.

Mark G. Herander: associate professor of economics, University of South Florida, August 1985.

R. Spence Hilton: senior economist, international research department, Federal Reserve Bank of New York, August 22, 1985.

Sandra C. Krieger: chief, securities department, Securities Analysis Division, Federal Reserve Bank of New York, August 22, 1985.

Cornelis Los: senior economist, domestic research department, Federal Reserve Bank of New York, August 22, 1985.

W. Douglas McMillan: associate professor of economics, Louisiana State University-Baton Rouge.

Khalid R. Mehtabdin: associate professor of economics, Niagara University, June 1985.

H. Dean Moberly: professor of economics, Auburn University-Montgomery, September 1, 1985.

Jan Palmer: associate professor, department of economics, Ohio University.

William M. Petersen: manager, banking studies department, Federal Reserve Bank of New York, July 19, 1985.

Rulon D. Pope: professor of economics, Brigham Young University, September 1985.

Roger L. Pupp: associate professor of economics, University of South Florida, August 1985.

Lawrence J. Radecki: research officer and senior economist, Research and Statistics Function, Federal Reserve Bank of New York, July 19, 1985.

Frank A. Scott, Jr.: associate professor, department of economics, University of Kentucky, July 1, 1985.

Steven M. Sheffrin: professor of economics, University of California-Davis, July 1, 1985.

Neil T. Skaggs: associate professor of economics, Illinois State University, August 16, 1985.

James I. Sturgeon: associate professor of economics, University of Missouri-Kansas City.

Eden S. H. Yu: professor of economics, Louisiana State University-Baton Rouge.

Michael A. Webb: associate professor of economics, University of Kentucky, July 1, 1985.

Thomas A. Zak: associate professor of economics, U.S. Naval Academy, August 16, 1985.

Administrative Appointments

Juan Amieva-Huerta: chairman, department of economics, Universidad Anahuac Mexico, September 1, 1985.

Don Bellante: chair, economics department, University of South Florida, July, 1985.

Roger E. Bolton: director, Center for Humanities and Social Sciences, Williams College, July 1, 1985.

John R. Finlay: chair, economics department, Wilfrid Laurier University, 1985–88.

Joseph A. Giacalone: dean, College of Business Administration, St. John's University, July 1, 1985.

Robinson G. Hollister: chairman, economics department, Swarthmore College, August 1985.

Werner Sichel: chair, department of economics, Western Michigan University, August 1985.

T. Norman Van Cott: chair, economics department, Ball State University, September 1, 1985.

Mahmood A. Zaidi: director, international program development, School of Management, University of Minnesota-Minneapolis, September 16, 1985

Appointments

Paul M. Anglin: instructor, department of economics, University of Kentucky, August 1985.

Klaus Becker, University of Kansas: instructor in economics, Ohio University, September 1, 1985.

Paul Beckerman: economist, external financing department, Developing Economies Division, Federal Reserve Bank of New York, August 26, 1985.

Ronald S. Blum, University of Wisconsin-Madison: visiting lecturer, Indiana University-Bloomington, August 1985.

Mark J. Brady: assistant professor of economics, Pitzer College, July 1, 1985.

Sunne Brandmeyer: lecturer of economics, University of South Florida, August 1985.

Scott J. Callan, Texas A&M University: assistant professor of economics, Clarkson University, September 1984.

Frank A. Camm, Jr., American Petroleum Institute: senior economist, Rand Corporation, September 1985.

Wendy L. Campbell, Cornell University: editorial and publications consultant, Research Findings in Print, Pittsford, NY, April 1, 1985.

Harvey Cutler, University of Washington: assistant professor, department of economics, Colorado State University-Fort Collins, August 20, 1985.

Paul DiLeo: economist, external financing department, Developing Economies Division, Federal Reserve Bank of New York, July 3, 1985.

Frederick Doolittle: economist, regional economics staff, monetary research department, Federal Reserve Bank of New York, June 3, 1985.

Lawrence B. Doxsey: director, economics and finance, Caliper Corporation, July 29, 1985.

B. Kelly Eakin, University of North Carolina-Chapel Hill: assistant professor of economics, University of Oregon, September 15, 1985.

Brian K. Edwards, University of California-San Diego: economist, Office of the Chief Economist, U.S. General Accounting Office, September 16, 1984.

A. Ramon Espinosa: economist, financial markets department, International Financial Markets Division, Federal Reserve Bank of New York, September 9, 1985.

Stefano Fenoaltea: visiting professor of economics, Swarthmore College, September 1985.

David W. Findlay, Purdue University: instructor of economics, Colby College, September 1, 1985.

Luke M. Froeb, Tulane University: economist, Litigation Economics Section, Antitrust Division, U.S. Department of Justice, August 1985.

John A. Galbraith, Treasury Board of Canada: economic advisor, Canada Deposit Insurance Corporation, October 1, 1985.

John E. Garen: assistant professor, department of economics, University of Kentucky, August 1985.

Charles L. Grim III, Purdue University: instructor of economics, Colby College, September 1, 1985.

A. Carter Hill, University of Georgia: professor of economics, Louisiana State University-Baton Rouge, August 1985.

Marianne T. Hill, Yale University: assistant professor of economics, University of Akron, September 1985.

A. Steven Holland: assistant professor, department of economics, University of Kentucky August 1985.

Kevin D. Hoover: assistant professor, department of economics, University of California-Davis, July 1, 1985.

Cheng Hsiao, University of Toronto: professor of economics, University of Southern California, July 1, 1985.

Wei-Chiao Huang, University of Connecticut: assistant professor of economics, Western Michigan University, August 1985.

Sadao Kanaya: instructor, department of economics, University of Kentucky, August 1985.

Bruce Kasman: economist, international research department, Industrial Economies Division, Federal Reserve Bank of New York, September 4, 1985.

Timothy C. Koeller: associate professor of economics, Stevens Institute of Technology, September 1, 1985.

Carl A. Kogut: assistant professor of economics, University of South Florida, August 1985.

L. A. Krause: instructor, department of economics, College of William and Mary, 1985–86.

W. D. Lastrapes: assistant professor of economics, Louisiana State University-Baton Rouge, August 1985.

Joseph Y. Lin: assistant professor of economics, Louisiana State University-Baton Rouge, August 1985.

James W. Mixon, Jr., University of North Carolina: visiting professor of economics, Brigham Young University, September 1985.

W. J. Moore, Miami University (Ohio): professor of economics, Louisiana State University-Baton Rouge, January 1985.

Michael J. Mueller, University of Oklahoma: assistant professor of economics, Clarkson University, September 1985.

Lynn Paquette: economist, financial markets department, Domestic Financial Markets Division, Federal Reserve Bank of New York, September 11, 1985.

Philip K. Porter: assistant professor of economics, University of South Florida, August 1985.

Bernard Rostker, Systems Research and Applications Corporation: director, Capabilities and Force Development Program, Arroyo Center, Rand Corporation, February 1985.

Subroto Roy, Virginia Polytechnic Institute and State University: visiting assistant professor of economics, Brigham Young University, September 1985.

George A. Rozanski, Harvard University: economist, Litigation Economics Section, Antitrust Division, U.S. Department of Justice, September 1985.

Todd Sandler, University of Wyoming: professor of economics, University of South Carolina, August 16, 1985.

John M. Santos, University of Illinois: instructor of economics, Colby College, September 1, 1985.

Morton O. Schapiro, Williams College: visiting associate professor, University of Southern California, September 1, 1985.

Jacques Silber, Bar-Ilan University, Israel: visiting associate professor, University of Southern California, September 1, 1985.

Sheryl R. Skolnick, Washington University: economist, Chicago Field Office, Antitrust Division, U.S. Department of Justice, July 1985.

Andrew Solocha, Michigan State University: assistant professor of economics, Clarkson University, July 1985.

Lee C. Spector, University of Iowa: assistant professor of economics, Ball State University, September 1, 1985.

Charles Steindel: economist, financial markets department, Domestic Financial Markets Division, Federal Reserve Bank of New York, August 19, 1985.

Jacob A. Stockfisch, American Petroleum Institute: senior economist, Rand Corporation, July 1985.

Joe A. Stone: professor of economics, University of Oregon, September 15, 1985.

Lawrence Sweet: economist, external financing department, Developing Economies Division, Federal Reserve Bank of New York, August 19, 1985.

Michael B. Tannen: visiting research professor, U.S. Naval Academy, August 16, 1985.

Mario Tello, University of Toronto: instructor, University of Southern California, September 1, 1985.

Katsuaki L. Terasawa, California Institute of Technology: senior economist, Rand Corporation, February 1985.

Jack E. Triplett, U.S. Bureau of Labor Statistics: chief economist, U.S. Bureau of Economic Analysis, July 1, 1985.

Geoffrey K. Turnbull: assistant professor of economics, Louisiana State University-Baton Rouge, August 1985.

John Veitch: assistant professor of economics, University of Southern California, September 1, 1984.

Eleanor T. Von Ende, University of Kansas: instructor in economics, Ohio State University, September 1, 1985.

Donald L. Westerfield, Southwestern Bell Telephone:

assistant professor of management and business administration, Graduate School, Webster University, August 1, 1985.

Larry E. Westphal, World Bank: professor of economics, Swarthmore College, September 1985.

Arnold Zellner, University of Chicago: visiting professor, University of Southern California, January 1, 1986.

Benjamin Zycher, California Institute of Technology: economist, Rand Corporation, January 1985.

Leave for Special Appointments

Mark S. Freeland, Health Care Financing Administration: visiting research economist, Institute for Health Policy Studies, University of California-San Francisco, September 1985–August 1986.

Clyde A. Haulman, College of William and Mary: Fulbright professor of economics, Wuhan University, People's Republic of China, 1985–86.

Resignations

Michael R. Baye, University of Kentucky, July 1, 1985.

Ali F. Darrat, University of Kentucky, July 1, 1985.

Timothy Hau, University of California-Davis, June 30, 1985.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

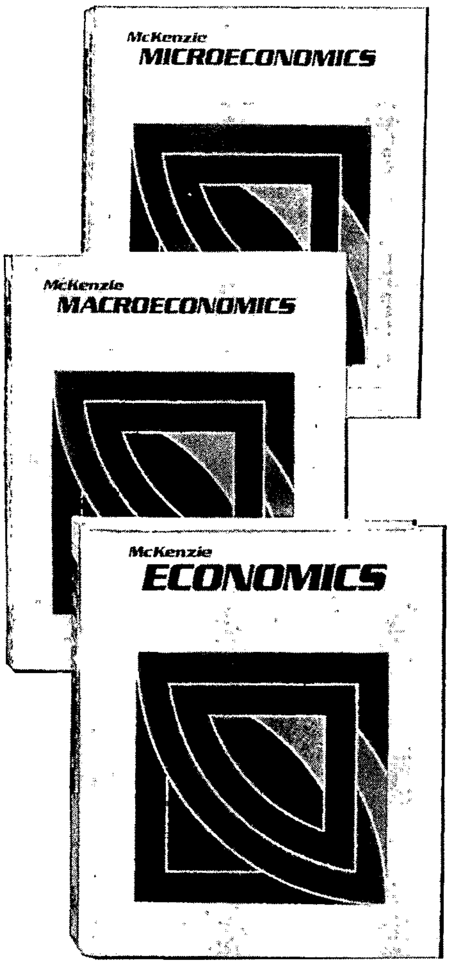
1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542–7067.



RICHARD McKENZIE

Clemson University

ECONOMICS
Complete Hardcover Edition: 700 pages

MACROECONOMICS: paper • 464 pages

MICROECONOMICS: paper • 576 pages

Study Guide • Instructor's Manual • Test Bank • Computerized Test Bank • GPA: Grade Performance Analyzer • Just published

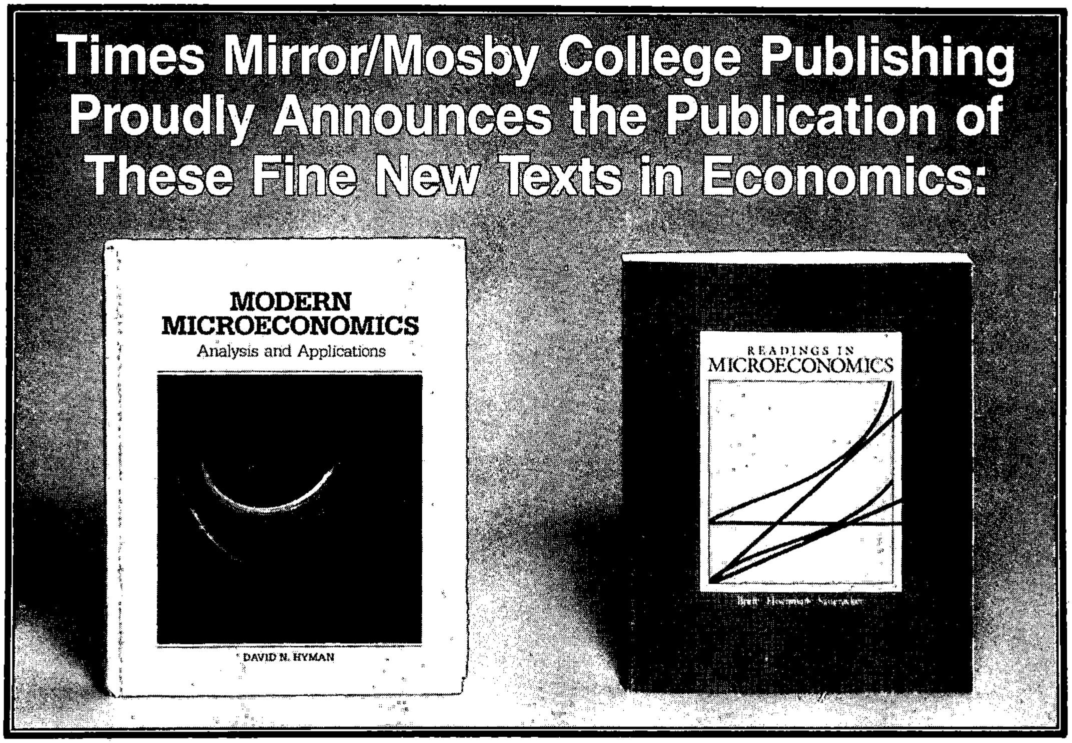
Combining experience in national policy-making and in teaching, Richard McKenzie offers a new textbook that presents economic principles in terms everyone can understand.

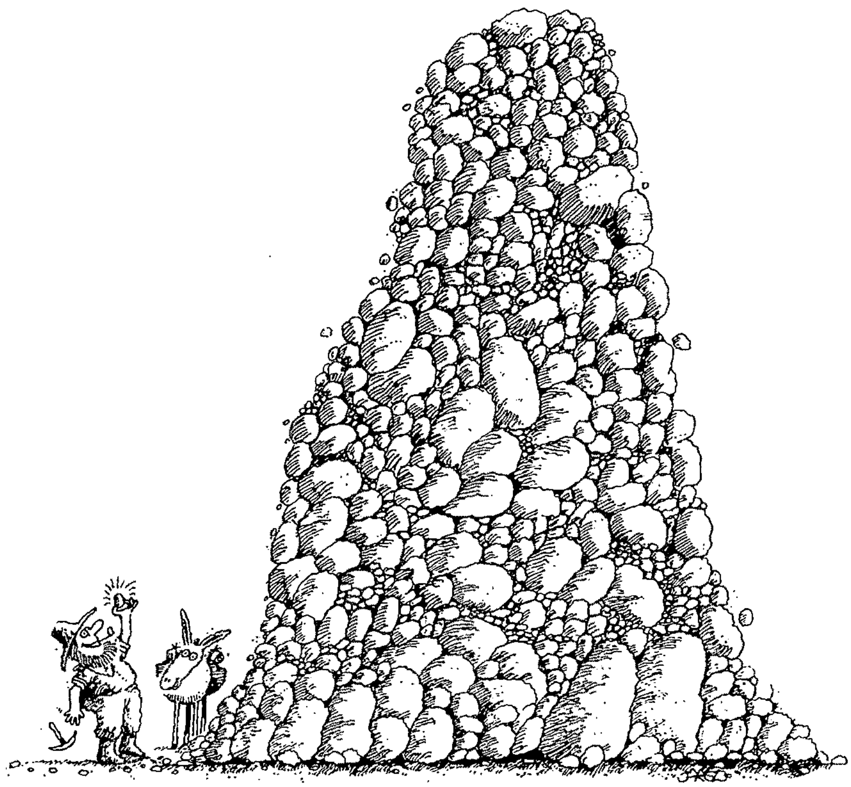
Boxed "Perspectives on Economics"—many written by specialists especially for this text—focus on today's critical issues.

For adoption consideration, request an examination package from your regional Houghton Mifflin office.

Houghton Mifflin Company

13400 Midway Rd., Dallas, TX 75244-5165
1900 S. Batavia Ave., Geneva, IL 60134
989 Lenox Dr., Lawrenceville, NJ 08648
777 California Ave., Palo Alto, CA 94304


Times Mirror/Mosby College Publishing
Proudly Announces the Publication of
These Fine New Texts in Economics:
MODERN MICROECONOMICS
Analysis and Applications
DAVID N. HYMAN
READINGS IN MICROECONOMICS



The American Economic Review

PAPERS AND PROCEEDINGS

OF THE

Ninety-Eighth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New York, New York, December 28–30, 1985

Program Arranged by Alice M. Rivlin

Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John

MAY 1986

THE AMERICAN ECONOMIC ASSOCIATION

•Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Survey*, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues (nonrefundable) for 1986, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- $37.50 if annual income is $30,000 or less;
- $45.00 if annual income is above $30,000, but no more than $40,000;
- $52.50 if annual income is above $40,000;
- $18.75 annually for registered students (three years only).

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), $105 a year; individuals, $70.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add $12.00 to the annual rates above to cover extra postage.

•Correspondence relating to the *Survey*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

THE AMERICAN ECONOMIC REVIEW

391
r 001

VOL. 76 NO. 2 MAY 1986

PAPERS AND PROCEEDINGS

OF THE

Ninety-Eighth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New York, New York

December 28–30, 1985

Program Arranged by Alice M. Rivlin

Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John

CONTENTS

PAPERS

PROCEEDINGS

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-eighth annual meetings of the American Economic Association. The *Proceedings* record the business activities of the Association in 1985; the annual membership meeting; and the March and December meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise eighty-two contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. We would like to take this opportunity to answer a number of commonly asked questions about the *Papers*.

Who chooses the authors? About a year in advance, the Association's President-elect, acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* is published annually in the Notes section of the December issue of the *AER*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing twenty-seven sessions, although a total of seventy-seven sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

Are discussants' comments published? There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect decided to publish no comments, given the difficulty and the invidious task of choosing. She has arranged instead that the names and affiliations of commentators be printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

What standards must the papers meet? The guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, very rigid deadlines must be met and there is not time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long, or does not satisfy other specifications.

This year, most authors cooperated very nicely. We thank them for making our lives easier. To those who failed to follow the guidelines, we suggest a reading of *Proverbs* 13:18.

HARVEY S. ROSEN
WILMA ST. JOHN

RICHARD T. ELY LECTURE

The Washington Economics Industry

By HERBERT STEIN*

When Alice Rivlin invited me to give the Ely Lecture I was, of course, honored, I was also puzzled. What could I talk about that most of the audience would not know as well as I, or better? I decided that the answer is the Washington economics industry. I have been working in, living in, and observing that industry for more than 47 years. Probably there are not over ten people who can say that. I was going to say that the other nine are in Palm Springs, but I see that Walter Salant is here. He nearly qualifies, so I will say that the other eight are in Palm Springs. Also, the 100th Anniversary of the Association seems an appropriate occasion for a little nostalgia. Moreover, Richard Ely, for whom this lecture series is named, had an interesting association with Washington economics. In fact, he seriously considered running for Congress in 1912 but was dissuaded by the president of the University of Wisconsin, who didn't think that was an appropriate job for a professor.

In the Ely Lecture last year, Sir Alec Cairncross referred to economics having become an industry. That is a term that economists do not usually apply to themselves. Probably few economists know that in the Standard Industrial Classification there is an industry called "Economic Research." It is a six-digit industry, 739210. We don't even know how many economists there are in the United States. The BLS has estimated that there were 110,000 economists in 1984. Every economist to whom I have mentioned this figure has been amazed at it. Of course, the BLS statistic comes from the *Current Population Survey*, in which people classify themselves by occupation. That introduces the possibility of bias. But why should anyone say that he is an economist if he is not?

If we start with the exceedingly rough estimate that there are about 100,000 economists in the country, variously employed, we can make an even rougher estimate of the income originating in the industry. According to the National Science Foundation, the average annual salary of economists in 1982 was $35,000. Probably each of these economists had on the average one secretary or other assistant not counted as an economist, with an average salary of $15,000. On the reasonable assumption that the amount of physical capital involved is negligible, that means an income originating in the industry of $50,000 per economist. Thus we have a $5 billion industry. That is about the same size as the motion picture industry, which is, like economics, involved in producing a combination of information and entertainment.

I propose, however, to concentrate on the Washington branch of the industry. And I shall rely heavily on my own observations and impressions rather than on statistical data. This is partly because the statistical data are meager. Also, my advantage in telling the story is that I was there, not that I had any special access to statistics, although I have collected some statistics for this occasion. You will recognize the inevitable subjectivity of my observations. Especially, what I report as the difference between 1938 and 1985 may be only the 47-year difference in the age of the observer and not the difference in the things being observed.

Let me assure you that I am not going to give a memoir of an ex-Chairman of the Council of Economic Advisers. That has become a boring literary genre, filled with episodes like this:

> "And then I said to the President, 'You should cut taxes,' or 'You should roll back steel prices,' or whatever."

*American Enterprise Institute, 1150 Seventeenth Street, NW, Washington, D.C. 20036.

"And the President said to me, 'You're a smart fellow, Tom,' or Dick, or whatever the name was, 'Not like that dumb Treasury Secretary I've got'."

"And that showed that although he had no formal training the President had a good intuitive understanding of economics and appreciation of economists, especially me...."

My story begins, then, in 1938, when I came to Washington. Thinking back at that now, I am surprised to realize that by then the executive branch of the federal government was already pretty well infiltrated by economists. Every agency for which it was at all relevant had a staff of economists. I worked for one of the smallest agencies, the Federal Deposit Insurance Corporation, and we had at least six economists.

According to a survey taken by the Bureau of Labor Statistics and the Civil Service Commission, there were 5,050 economists employed by the federal government at the end of 1938, not including home economists. Probably the most realistic figure for the present is the 16,000 economists in the federal government estimated by the National Science Foundation for 1982. That would mean that the number of economists had roughly tripled, which is about what has happened to the total federal civilian work force. Of course, not all of those government economists worked in Washington and its environs, but about two-thirds of them do now and probably the same proportion did then.

The number of economists in the federal government had increased greatly in the years before 1938. In a paper given at the meeting of the American Economic Association in December 1936, Leonard White, Chairman of the Civil Service Commission, estimated that the size of the federal economics and statistical services had doubled within the previous two years. He also referred to a study estimating that in 1931 there had been 742 positions in econmics and statistics, not including home economists, White said: "This may or may not correspond to the accepted definition of a professional economist. The figure doubtless represents a maximum if we are thinking of 'genuine' economists." White put the word "genuine" in quotation marks. What he meant by that, or by the accepted definition of professional economists, I do not know. It is not a distinction I intend to make.

Pushing the history back before 1930 in statistical terms is difficult. In 1896, according to White, there had been 87 people in the government with the title of "statistician." But there was only one with the word "economics" in his title, and he was an "economic ornithologist." That only shows how unreliable these classifications are. Many economists had worked for the federal government, not only before 1930 but also before 1900. Francis Amasa Walker, the first president of this Association, had been Director of the Census in 1880 and 1890. In fact, half of the men who were presidents of the AEA during the first fifty years of its existence had some experience working for the federal government. The Census Bureau and the Interstate Commerce Commission were favorite places.

Richard Ely did not work for the federal government, although he held several positions for the state of Wisconsin in Madison. He was, however, much interested in the employment of economists in the federal government and seems to have been uncommonly diligent in promoting his students and colleagues for federal jobs. He apparently thought that there was a given number of University of Wisconsin slots in the federal service. (The following facts about Ely and his contemporaries of the years before 1920 are derived from a fascinating dissertation "Professors and Public Service, 1885–1925" by David M. Grossman.)

When Walter Willcox, later to be president of the AEA, was appointed as chief statistician of the Census, in 1899, he wrote to Ely as follows:

"It may conduce to a better understanding of the situation if I explain that the selection of a college teacher for one of the positions was not intended as a compliment to the University he represented...."

Ely and others of his time did not regard Washington as a useful place for a career; it was a place for temporary exposure to the real world before returning to the

serious business of the University. In 1906 Ely wrote to John R. Commons about government service:

"I believe that very generally it is felt that two or three years of that sort of work is about as much as a man can safely undertake."

These economists before World War I were generally involved in technical positions in Washington, often as statisticians. During World War I some economists came to Washington in positions with more policy responsibility—Frank W. Taussing at the Price Fixing Commission, Wesley Mitchell with the War Industries Board, and Edwin Gay with the War Shipping Board. Wars have played a major part in the development of the Washington economics industries.

These early notables, although they spent some time in Washington, were not what we would call today Washington economists. That is, their careers were not made in Washington. An interesting case is that of Henry Carter Adams, who, when he accepted a post with the Interstate Commerce Commission, insisted on the creation of a field office in Ann Arbor which would allow him to devote most of his time and energy to the university.

Let me now return to 1938. Opportunities for the employment of economists in the academic world or in business were not good. Fellowships for remaining in graduate student status for years and years were not available. Economists turned to the federal government because that was where the jobs were. Many came to Washington expecting to earn a living while they finished their theses, which some did and some found, many years later, they had not done. To some extent the increase of employment of economists in the federal government was supply-driven. The federal government was in the business of providing jobs for all kinds of people—artists, actors, writers, for example. Some of the employment of economists came about in this permissive atmosphere.

But also the economists, mostly young, who came to the federal government 50 years ago found it an interesting, satisfying place to work—beyond the satisfaction of having a job and an income. It is not that we thought ourselves to play important roles in great revolutionary decisions. When I came to study the period later, to write *The Fiscal Revolution in America*, I was surprised at how little I had known of what was going on although I was there. That was not a time when every whisper of a hint of a thought about economic policy was reported in the Washington press and mulled over at lunch by every economist in town. But there was a certain freshness and interest in the tasks that we were doing, even though they were small tasks. We were working in new agencies or on new assignments of old agencies, and we were bringing to many of these tasks for the first time the economics we had just learned. Much of what we were doing would now be regarded as drudgery. Work that took me months at the FDIC would now be done by a computer in minutes. But at least I thought I understood why I was doing it.

One might think that although economists were spread widely through the executive branch by 1938 they were not as influential as they later became. But I am not sure that was true. People like Harry White at Treasury, Isadore Lubin at Labor, and Herbert Feis at State were important. In fact, they were probably more important than any subsequent economist in those departments until an economist become secretary in each of them. And, but for the accident that George Shultz is an economist, it might still be true that those people in the 1930's were the most influential economists those departments ever had.

There was no president's Council of Economic Advisers, of course. But the president did talk to some economists—Lauchlin Currie, Isadore Lubin, Leon Henderson, for example. Whether their influence was as great as that of later Councils of Economic Advisers over their presidents is uncertain. Probably the difference was not dramatic. Roosevelt got all the advice of economists he wanted, which is about as much as any president gets.

In one respect, at least, the late 1930's was the golden age for government economists. That was in the development of economic statistics. This was the period when scientific sampling was introduced into federal statisti-

cal methods. The chief product of that was the *Current Population Survey*, which yields the employment and unemployment figures and much else. It is of some interest that the *Current Population Survey* was initiated by the WPA, which existed primarily to provide jobs. Also in this period the official national income statistics were initiated and developed most of the way to their present condition. A large part of the statistics upon which economists rely heavily today originated in the years 1935–40.

These years, just before World War II, were not a period of sharp division among Washington economists on grounds of theory or ideology. We were all Keynesians in a loose, eclectic, pragmatic sense. We thought that the immediate economic problem was inadequacy of total demand and that the long-run problem was instability of total demand. We thought that a flexible fiscal policy was a useful and possibly essential means for correcting these problems. By 1938 interest in structural or micro issues, the kind of interest that had given rise to the National Recovery Administration, had faded among economists. There was a view in Washington that industrial monopoly was a major source of our economic problems, but that view was largely confined to lawyers. The attention of economists was focused on macroeconomics. It was not until Alvin Hansen revealed his version of Keynesian economics in 1941, and Henry Simons replied to it in 1942, that I realized that I was not the same kind of Keynesian as everyone else.

A little nest of dissent existed at the Brookings Institution. Brookings was then the only nongovernment research institute, the term "think-tank" not yet having been invented. It was a product of World War I, when Robert Brookings, a St. Louis businessman serving in wartime economic agencies, became convinced of the need to apply social science to government. I worked for a man who had been a fellow at Brookings and we went there frequently for lunch. The stalwarts of the institution who only a few years earlier had been in the front line of economic discussion, now seemed to talk an archaic language and to be completely out of the picture. They worried about the federal debt, and were either 10 years behind the times or 40 years ahead.

I do not think of the Washington economists of 1938 as being identifiably Democrats or Republicans, as would be true of many today. Perhaps that is because they were all Democrats. But I don't think that is the whole explanation. Certainly the upper level of government economists were devoted to Roosevelt and the New Deal. But that was not the same as being a Democrat. For many economists, serving in the government during the New Deal was a nonpartisan activity, like serving during a war.

World War II was watershed in the history of Washington economics. World War I had established the fact that economists could be useful in the operation of a war economy. But the infusion of economists was much greater in World War II. Of course, the second war was much bigger and longer and had much more extensive economic controls. But also, the second war came to a Washington that already had a large number of government economists, they were naturally drawn to the early stages of economic planning for the war, and they naturally recruited other economists as the control activities expanded.

The most important contribution of economists to the war effort was not in the management of the price and production controls, which was probably as well done by lawyers, accountants, businessmen, and engineers. It came in the big decisions about the overall size of the effort. The economist who made this contribution was Robert Nathan, who turns out in my story to be one of the outstanding figures in five decades of Washington economics. He had been one of the pioneers in the development of the national income accounts at the Department of Commerce. After the war he would pioneer the movement of Washington economists into the world of profit by establishing one of the first private consulting firms. In the early days of the war, before the United States entered, there was substantial disagreement about how much military build-up the economy could afford. Using the national income statistics, Nathan made an effective argument that we could afford a much bigger build-up than the War and Navy Departments were planning. After Pearl Harbor the

situation was reversed. The services wanted everything. Nathan by then was director of the Planning Division of the War Production Board. He showed that the combined requirements could not be met and that the attempt to meet them all would result in unbalanced forces, uncompleted weapons systems, and wasteful production.

Macroeconomists can feel confident in wartime, because in wartime they deal with large numbers—large enough to override the noise in the data and the conditionality of the analysis. We may not predict very well the consequences of the difference between federal spending of 20 or 25 percent of *GNP*, or of a deficit of 2 or 3 percent of *GNP*. But we can give a useful, if rough, estimate of the consequences of raising federal spending from 10 to 50 percent of *GNP*, or of raising the deficit from 3 to 25 percent of *GNP*.

The war years were the time of the closest connection between Washington economics and the rest of the economic industry. One evidence of that is the frequency of publications in economic journals by government economists. George Stigler once made a tabulation of the sources of articles in four general economics journals at 10-year intervals from 1882–83 to 1962–63. With the help of my secretary, I have repeated that for 1972–73 and 1982–83. In the years 1942–43, 17 percent of the articles came from economists in government. The average of all the other years was only $5\frac{1}{2}$ percent and the highest figure side from the war years was 9.4 percent in 1912–13. It was not only that more economists were in government during the war. It was also that a large proportion of the articles in the journals were related to war or postwar problems. The program of the 1944 annual meetings of the American Economic Association was almost entirely filled with papers on such subjects. As a further contribution of the profession to the war effort, the meetings were cancelled at the request of the Office of Defense Transportation in order to save transportation.

By the end of the war it was clear that a large role for the federal government in the economy was here to stay. The 1930's fight about the New Deal was over. For one thing, scores of businessmen, the chief recalcitrants about the New Deal, had spent the war in Washington and learned that they could live with it. And it was also clear that the language in which this activity would be conducted was economics. It was almost as if someone had suddenly decreed that the language of the government would be Latin. There would be a great demand for people who could speak Latin. So there was a great demand in Washington for people who could speak economics. There was also a large supply of them, who had come for the war and didn't want to go home again.

The growth of the Washington economics industry proceeds largely by competition and emulation. Washington is an arena of competition—the White House against the departments, the executive branch against the Congress, the regulated against the regulators. If one team has economists, the other team must respond by having economists. Much, of course not all, of what has happened in the past 40 years can be explained by this process.

I shall not try to unravel the tangled web of statistics about the size and growth of the Washington economics industry. About all one can say with confidence is that there are lots of economists in Washington, that their numbers have increased substantially, and that the growth of Washington economics outside the government since before the war has been much greater than the growth inside the government. For those who insist on numbers, I will offer the following guesses. I think that there are about 15,000 economists in the Washington area, excluding academics, compared to about 3,000 in 1938. Of those 15,000, about 11,000 are in the federal government, including about 500 in the legislative branch, where there were probably none in 1938. The 4,000 outside the federal government are divided in about equal thirds among international institutions—the IMF, the World Bank, and their smaller brothers—nonprofit research institutions, think tanks—and representatives of the private profit sector, including consultancies. Almost all of this is new since World War II. In 1938 there were only eight economists listed in the yellow pages of the Washington phone book. I should add to the category of the Washington economics industry a handful of people who will be forever identified with Washing-

ton, whether they live in Minneapolis, Ann Arbor, St. Louis, or wherever.

One real statistic I cannot forebear to mention. According to the National Science Foundation, the number of economists in the federal government increased by about 60 percent from 1980 to 1983. I wonder whether President Reagan knows that.

If my estimate is anywhere near correct, there are about two-thirds as many economists in the Washington economics industry as on all the college and university faculties in the country, and about three-fourths as many as belong to the AEA. Also, clearly the Washington economics industry is much more than the fifteen or so people who appear on TV all the time. They are to the Washington economics industry what Lee Iacocca is to the automobile industry.

As far as the executive branch is concerned, the main development of the postwar years is not the increase in the number of economists. The increase has apparently been large, but that is a continuation of the trend noticed before the war. The main development in the executive branch was the change in the level of economists. Starting with the Council of Economic Advisers, we got an increasing number of economists of cabinet or sub-cabinet rank, appointed by the president, confirmed by the Senate and serving at the president's pleasure. Even the leading members of the Roosevelt Brain Trust did not have such status. That had a number of implications that I will point out later.

What are all of these economists doing in the Washington economics industry? Broadly, their functions can be divided into three—research, advocacy, and decision making. This does not mean that the individual economists can be segregated into these three boxes. Many do two or three of these things, although I suppose most Washington economists could be identified as being primarily engaged in one.

I shall have little to say about the advocacy function. It comes in two parts—micro, that is, advocacy before a limited audience, like a regulatory agency or a congressional committee, for a specific purpose, like a rate increase or a desired tax treatment—and macro advocacy, that is, trying to appeal to a broader audience, ultimately the public. I have little experience with micro advocacy, except to observe that its practitioners eat lunch at the expensive K Street restaurants and smoke big cigars. Most of my life in Washington was spent in a combination of research and macro advocacy, at the Committee for Economic Development. This function is commonly misunderstood, even by those who participate in it, and perhaps especially by them. They believe, and tell their supporters, that they are having a great influence on public policy by talking to decision makers in Washington and appearing before congressional committees. The fallacy of that became clear to me when I was later in the government. The self-appointed outside advisers to government do not function quickly enough, or with enough information, to have much direct influence on the government decision-making process. Where they do have an influence, and I believe it is a real influence, is in gradually affecting the climate of public opinion within which government officials feel they have to operate.

According to the National Science Foundation, 80 percent of all federal government economists in 1982 reported that their primary work activity was research and development, the management of *R&D* and reporting and statistical activity. Two questions come to my mind about this. The first is whether government economists, presumably engaged in research of some kind, are keeping up with developments in the academic side of the industry. I am led to this question by observing that when I came to Washington, I could read and reasonably understand almost everything in the economic journals. Today I can read hardly any of it. Also, I observe that although government economists are about 15 percent of all economists and about two-thirds as numerous as the economics teachers, they contribute only 3 or 4 percent of the articles in economic journals. This is a smaller percentage than at the turn of the century.

Further reflection and observation lead to more reassurance. When I recall the economists with whom I worked at the Council of Economic Advisers and solicit the experience of my colleagues at the American Enterprise Institute, I am persuaded that while economics has left me behind it has not left the

Washington economics community behind. I often think of the occasion at the Council of Economic Advisers in 1969 when we were asked what the effect of repealing the investment tax credit would be on economic growth. One of the young members of the staff, drawing on the latest journal articles, produced the answer overnight, to two decimal places. It may not have been the right answer, but it was the best economics had to offer.

As for why government economists do not write more for economic journals, I am reminded of the answer the woman gave when asked why her 20-year-old son was carried from the car to the apartment house by the chauffeur. "Of course he can walk, but thank God he doesn't have to."

The second question about the research activities of the Washington economists is whether they are adequately serving their function of providing the rest of the profession with the data, mainly statistical, that it needs. If, for example, you compare the statistical appendices to the CEA's *Economic Reports* for 1960 and 1985, you will find some increase of detail in the later year, but hardly any new subjects covered except for the poverty and productivity statistics. Perhaps that is not a sufficient sample, but still I think it is clear that the pace of advance in supplying new data has slowed down from the 1930's and the first postwar decade. It may be that we have measured everything that could or should be measured, but I doubt that. I have the fear that unless some new data are turned up soon, the econometricians will be unemployed for lack of inputs, having exhaustively analyzed all the existing information. Of course, Washington does provide occasional revisions of the old data, which provides some new work to do.

Let me turn now to the part of the Washington economics industry that is in or near the policymaking process, or aspires to be. That is certainly a small part of the industry, may not be the most important but is certainly the most conspicuous. Three things strike me about this part of the industry, especially as compared with my impression of the situation in the 1930's. These things are all related. The participants are more political, they are more divided, in talk although not in action, into different schools of economics, and they get much more publicity.

I have said elsewhere that the Employment Act of 1946 may or may not have succeeded in its goal of introducing economics into politics, but it certainly succeeded in introducing politics into economics. This would have happened anyway, in time, even if there had been no Council of Economic Advisers. Once presidents and cabinet members get to appointing economists who will be close to them, speak for them, and on occasion act for them, they look naturally for sympathy and loyalty, and they are likely to take party affiliation as evidence of that. Even where economists are not chosen on that basis, they tend to acquire a feeling of membership in the political team if they have been treated as members of it, and few who have had that experience ever regain their virginity. And economists who aspire to such positions, or who only enjoy the excitement of involvement in the political process, take on political coloration and offer themselves to candidates who all now want economists on their campaign planes along with the make-up men and the TV producers.

So we have developed a cadre of economists who are identifiably Republican or Democrat. Even beyond that we are developing coteries that are particular kinds of Republicans or Democrats—Kemp followers or Bush beaters, or what not.

Connected with their partisan identification is an increasing tendency to differentiate the product that economists have to offer, as far as talk is concerned. If you are the adviser to a Republican candidate or incumbent, you tend even in your own mind to magnify your differences from the advisers of the Democrats. So you get little schools of economists who cling together, help each other, and talk as if they had serious ideological or theoretical differences with the others, about incomes policy, or monetary rules, or the natural rate of unemployment or other things. But the relations among these political, ideological schools have been amicable, with some exceptions. The participants seem to recognize that although they are on different teams, they are in the same game and that they need each other.

Moreover, when these people get into positions of responsibility, their differences fade. I think that a study of the recommendations and reports of the Council of Economic Advisers during the 28 years from 1953 to 1981 —16 Republican years and 12 Democratic—would reveal little durable difference among them. Each team comes into office with a grand manifesto emphasizing its fundamental differences from its predecessors. Within a year or two, that is all submerged in the attempt to wrestle with the same problems with the same inadequate instruments and knowledge. The Reagan team of economists entered office with exceptionally large differences from all of their predecessors, Republican as well as Democratic. But by now the sharp edges of those differences have worn away. The most devoted supporters of what was peculiarly Reagan economics have left the team and accuse those who remain of infidelity. To some on the outside this tendency towards an eclectic, mainstream position looks like weakness and political compromise. In my opinion, it is compromise all right, but with reality, with the reality of the limitations of economics and with the intractability of the economy. Just as there are no atheists in foxholes, there are no dogmatists at the Council of Economic Advisers, and for similar reasons. The risks are too great.

The greatly increased publicity of Washington economists in the popular media since the end of the war is connected with their politicization. The media are interested in economics mainly as a branch of politics. What they want to know about the tax bill is not what it means for the economy, but what it means for President Reagan or Congressman Rostenkowski. They look to economics as a window for observing politics and they are attracted to economists with political connections.

As an attempt to measure the publicity given to Washington economists, with the help of my wife, my only research assistant, I have counted up the number of times the *New York Times* index referred to what I considered the most conspicuous Washington economists at intervals beginning in 1938. In 1938, there were 11, or 23 if one includes Leon Henderson who was an economist not functioning as an economist. In 1948 and in 1953, there were 34. In 1984, there were 74. The year 1972 was off the trend with 132, because that was a year of great excitement about price controls as well as about a presidential election.

I have also counted the number of appearances by economists on Meet the Press in the 37 years of its weekly programming. There have been 57 such appearances, far more than for any other group of people other than full-time politicians. (I exclude people like George Shultz and James Schlesinger who are economists but not being interviewed as economists.) Almost all of these appearances were by Washington economists, except for the occasional appearance by a newly crowned Nobel laureate. And when the little red light went on and the camera was rolling, the Nobel laureates all sounded like Washington economists anyway —just as partisan and just as willing to give easy answers to hard questions. Twelve of these appearances were by Walter Heller who, although he no longer lives in Washington, will always be the Washington economist par excellence. Heller even appeared more often than King Hussein. The only three people who had more appearances than he were Henry Cabot Lodge, Hubert Humphrey, and Nelson Rockefeller—all of whom had been vice presidents or candidates for the vice presidency.

This increased public exposure is gratifying to a degree, but also worrisome. It is largely confined to the more political members of our profession, who are not always the most candid or best informed. These people tend to emphasize controversy. Frequently the TV producer will call up to ask what answers you would give to certain questions if invited to appear on the program, and this usually sounds like the try-out for a combat. Frank Knight liked to repeat the saying that the trouble was not what people didn't know, but what they knew that wasn't so. I am not sure whether our public appearances are reducing what people don't know more than they are increasing what people know that isn't so. But I am reassured when I consider the alternatives. The TV time must be filled, and the FCC requires that a certain amount of it be filled with talk.

If someone is going to talk about economics on TV, it is probably better done by economists than by politicians, columnists, sociologists, or clergymen—who seem the most likely alternatives.

Probably that applies to much of what the Washington economics industry does. We may not do very well what we do, but if it has to be done it is better done by us than by the alternatives.

I observe that most articles in the *American Economic Review* have a conclusion. I have always supposed that this was because the body of the article is so impenetrable that one last-gasp effort had to be made to explain what it all means. And I suppose I should have a conclusion evaluating the Washington economics industry. But how does an economist evaluate an industry? We are not Ralph Naders, asking whether the industry causes cancer or highway fatalities, or even whether it is good or true. We may ask whether the industry is profitable, but that is not measurable for most of the Washington economics industry.

Probably the best standard for evaluating this industry is the quality of life experienced by the people who participate in it. By this standard the statistics are ambiguous. The average salary of a government economist was almost 12 times as high in 1983 as in 1938. Deflated by the Consumer Price Index, the real increase was almost 70 percent. But, in the same period, real annual wages and salaries per full-time equivalent employee in the economy as a whole increased by 120 percent.

Even this computed 70 percent increase in the real incomes of government economists does not seem realistic to me. I feel like the housewives I am always meeting who tell me that their cost of living has gone up more than the CPI. But when we came to Washington in 1938 we could rent an apartment four blocks from the White House for $50 a month and hire a maid for twenty-five cents an hour. There is no comparable rent today and the maid's wage would be at least twenty times higher. As I think of the living standards of people in comparable positions today and in 1938, it seems to me that the main differences are that today's economists have color televisions and trips to Europe, with which they are getting bored, and less domestic service, if any.

But still, the Washington economist leads a comfortable life in the dimensions measured by income statistics. The average salary of a Washington economist is about twice as high as the salary of the average American worker. It is a little less than the salary of the average Washington lawyer, but not enough less to be irritating. And there is more to life than can be measured by income statistics. The Washington economist is likely to have a job with some intellectual challenge. He faces a certain amount of competitiveness or rivalry in his professional life, enough to make the game interesting, but the game is played with civility—more civility than I understand is found on many college campuses. This civility is, I believe, derived from the style of politicians, who may be ruthless, but are also in the business of being nice to people. The Washington economist, if he has even a little imagination, gets an occasional thrill out of feeling that he lives in the shadow of history. He gets a kick out of being connected, even in a small way, with events that are part of the daily news not only of the city, but also of the nation and the world. Of course, there is a certain repetitiveness in the news—the new tax simplification plan, the new plan for balancing the budget, the new turn of monetary policy. After 47 years, one has the sense of having gone around this track many times. But, after 47 years, that happens everywhere. Even the pilot of the space shuttle must say, after the 47th orbit, "Gee, here comes that old Great Wall of China again." But for the first 45 orbits it is all very exciting.

I hope you will not think it too prosaic or even belittling that I have summed up my observations on the Washington economics industry by appraising the lives of those who work in it, instead of by judging how the industry has changed the world for good or ill. Even though we are economists we like to think that we have not only a price, but also a value, which is presumably greater. I do not want to deny that there may be another standard for judging the industry. But that will have to wait for another occasion and another judge.

Unnatural Value: or Art Investment as Floating Crap Game

By WILLIAM J. BAUMOL*

I shall suggest on the basis of a priori considerations and several centuries of price data that in the market for the visual arts, particularly the works of noted creators who are no longer living, there may exist no equilibrium level, so that the prices of such art objects may be strictly *un*natural in the classical sense. Their prices can float more or less aimlessly and their unpredictable oscillations are apt to be the exacerbated by the activities of those who treat such art objects as "investments," and who, according to the data, earn a real rate of return very close to zero on the average. If the art marketing process really is inherently rudderless, the imperfection of the available information on prices and transactions does not matter in the sense that better information about the behavior of the market really would not help anyone to make decisions more effectively.

I. Supply Response: The Pricing Anchor for Manufactures

The art market contrasts sharply with those for manufactured products, such as steel bolts or ball bearings, in terms of determinancy of equilibrium price level. There the key to equilibration is responsiveness of supply. If, for example, a manufactured product's current market price happens to be well above its equilibrium level, as the text books tell us, capital will flow into the production of the overpriced commodity, its output will be increased and its price driven downward. Thus the equilibrium price comes equipped with a powerful magnet capable of attracting actual market prices to it.

It is this mechanism that imparts value to pertinent information, for data on costs, on the nature of demand, and on the cost of capital are of value primarily because they help the observer to evaluate the equilibrium price, which is of practical interest *only* if there exist reliable forces pulling the actual prices in its direction.

II. The Unanchored Prices of Noted Works of Art

We may well suspect, in contrast with the manufacturing case, that the equilibration process will be considerably weakened in a market where elasticity of supply is absolutely zero, as it is in the market for the noted works of noted but deceased artists (an occasional intrusion of forgeries aside).[1] One may even surmise that, as in stock prices, the market values of such works of art will exhibit random behavior.

Indeed, there are several distinctions between the workings of the securities and arts markets, all of which suggest that an equilibration mechanism is likely to be more feeble in the latter.

First, the inventory of a particular stock is made up of a large number of homogeneous securities, all perfect substitutes for one

[†]*Discussants*: William S. Hendon, *Journal of Cultural Economics*; Harold Horowitz, National Endowment for the Arts; Virginia Lee Owen, Illinois State University.

*Princeton University, 108 Dickinson Hall, Princeton, NJ 08542, and New York University. I should like to express my deep gratitude to the C.V. Starr Center for Applied Economics for its generous support of the research underlying this paper. I am heavily indebted to Michael Goldberg for his analysis of the price data, and to Michael Montias for his very valuable comments, though some differences in views remain between us.

[1]I deal here with noted works by noted artists because the markets for the products of what are considered minor schools work very differently. As Montias has pointed out, a sudden rise in the popularity of such a group can elicit a flow of their works from attics and basements, thereby rapidly expanding their available supply.

another. Widely known paintings and sculptures are unique, and even two works on the same theme by a given artist are imperfect substitutes.

Second, a given stock is held by many individuals who are potentially independent traders on the near perfectly competitive stock market. The owner of a Cranach or a Caravaggio holds what may be interpreted as a monopoly on that work of art.

Third, transactions in a given stock take place frequently, indeed, almost continuously. The resale of a given art object may not even occur once in a century.

Fourth, the price at which a stock is exchanged is, generally, public information. The price at which an art work is acquired is frequently known only to the parties immediately involved. While, as I will argue, the availability of such information is not so helpful as is sometimes believed, it surely is unlikely to impede equilibration.

Finally, in the case of a stock we know, at least in principle, what its "true" (equilibrium) price should be—it is the stock's pro rata share of the discounted present value of the company's expected stream of future earnings. But, for a work of art, who would dare to claim to know the true equilibrium price? Distorting Oscar Wilde to my purposes, even those critics who claim to know the value of everything may know the true price of nothing.

In these circumstances it seems implausible that art markets possess anything like long-run equilibrium prices, let alone that there exist reliable forces that drive market prices toward them.

III. On the Economic Value of Art Market Information

Those economists who helped to achieve it are proud of their role in the unbundling of the services of stock brokers, in good part because, as a result, the securities purchaser is no longer required to pay for research which most economists consider to be useless to the investor. If stock prices do indeed approximate random walks, as the evidence strongly indicates, then there is little that information can do to improve estimates of future prices, the key forecast for the purchaser of stocks.

But, if art prices are no more orderly than the prices of stocks, and perhaps even considerably less so, how can data on past activity in the art market conceivably serve as a portent for the future? If stock market research is worthless for the stock market investor; if the stock purchaser can select as well by throwing darts at the financial pages as by following the advice of professional analysts (see, for example, Burton Malkiel, 1973), how much better off can the investor in art hope to emerge with the aid of similar data on art sales with all their warts and blemishes, or even with the help of someone who conducts some sort of "analysis" of those data, perhaps on the lines of the fundamental or technical approaches fashionable among stock market analysts?

IV. Some Data and their Rate of Return Implications

While data on the art market are woefully incomplete and even those that are available are not easy to come by, there exists a remarkable source which permits analysis going beyond anything I have encountered in the literature. In one book of a three-volume set, Gerald Reitlinger (1961) provides an extensive compendium of the sales of art works by "...the best known painters of the world,"[2] extending over more than five centuries. A price is given for each reported sale, which seems to include every transaction involving the work of a painter on Reitlinger's list for which price data are known to be recorded. As the author describes it, "unless otherwise stated, the items refer to London sales. Until 1920 or thereabouts this means with few exceptions sales at Christie's" (p. 242).

[2]It is a noteworthy comment on the haphazard fluctuation of tastes that in the same passage in which Reitlinger ponders on the curiously long period during which Vermeer was ignored, he justifies his inclusion of Turner by the fact that he was a "...Monarch...in the salesroom of [his] day and a very curious chapter in the history of taste, which is so often the history of bad taste" (p. 241).

The art market simply does not provide the continuous data or even the continuous transactions that would be required for a systematic analysis of sophisticated issues such as a random walk hypothesis. However, analysis of simpler issues remains possible. Specifically, I will turn now to examination of the rate of return on investment in art.

Of the thousands of sales recorded between pages 241 and 506 of Reitlinger's book, there are a substantial number of cases in which a given work of art was resold two times and more during a 300-year period. We compiled a complete list of such multiple sales and their prices, and sought to determine what range of rates of return the investor could have hoped for during this period.

Specifically, the following procedure was employed: from the complete list of multiple sales we eliminated all cases in which an interval of less than 20 years intervened between the sales. Approximately 25 listings involved some inconsistencies and were eliminated. In another 25 or so, there were no firm price figures but only word of mouth financial information, and they too were eliminated.

This left us with a total of 640 transactions extending from 1652 to 1961. The reported prices were then deflated by a price index to transform them into pounds of constant purchasing power. For the years 1652 to 1952, the E. H. Phelps-Brown and Sheila Hopkins (1956) index of the prices of consumables was employed. For the period 1955–61, deflation was carried out using the International Monetary Fund Consumer Price Index (1979). The two indices, of course, do not match perfectly but permit a workable deflation procedure.

Finally, from these deflated figures, rate of return figures were calculated for each painting for the period between adjacent transactions. These were calculated from the standard continuous compounding formula $y_t = y_0 e^{r(t-t_0)}$. From these a set of measures of central tendency, that is, the mean, median, standard deviation, etc. were determined and a histogram of the observations was prepared. Let us, then, see what these showed.

V. Results

As a standard of reference it should be noted that, apart from the time of the Napoleonic wars and a few other episodes that were relatively brief, the rate of inflation during the period that encompasses our data was extraordinarily low by current standards. Indeed, by and large the nineteenth century can be characterized as a period of deflation. Over the 300-year span containing our cases, the Phelps-Brown and Hopkins price index rose at an average rate less than 0.7 percent per year. At the same time, according to Sidney Homer (1977), the rate of interest on the safest securities of the British government ranged from a high of some 6 percent near 1800 during the Napoleonic wars, to a low of about 2.25 percent during the Victorian "great depression" of the 1890's in Britain. These include the famous "consols" which have no redemption date and which, literature recounts, were the mainstay of Victorian widows or surviving spinster daughters from financially comfortable families. Probably about 3.25 percent was a representative nominal rate of return for the period, providing a real return of, perhaps, 2.5 percent.

Now it should be recognized that ownership of a painting is a risky affair, aside from whatever financial uncertainty may be involved. A painting can be stolen or destroyed in a fire. English collectors after the restoration were spared the risk caused by wars and revolution (though the affair of the '45 glamorized by "Bonnie Prince Charlie" may have seemed rather a near thing at the time). Yet, London had undergone its great fire in 1666 which left, perhaps, one-fifth of the walled city intact, and organized firefighting techniques only arose well into the nineteenth century. The implication is that whatever the apparent rate of return the ownership of a painting yields, a substantial risk premium must be deducted from the figure to get at the true underlying rate of return.

In addition, the sales commissions charged by the sales agent should of course be subtracted from an art work's resale price in

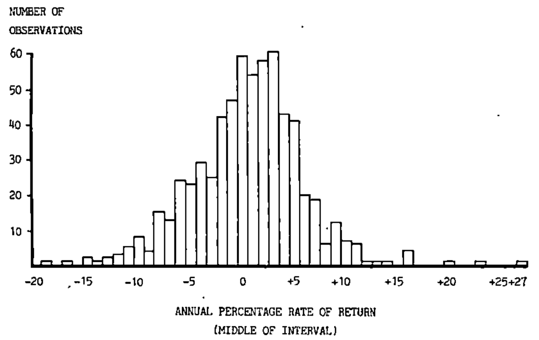


FIGURE 1

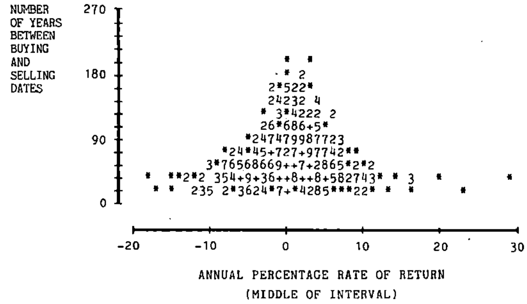


FIGURE 2

Note: Each point is plotted with an asterisk. When more than one point falls on the same plotting position, a count of the number of points falling there is given. When more than nine points fall on the same plotting position, the + symbol is given.

order to evaluate the true rate of return to the investor. Having no data on such selling fees in earlier centuries, we made no attempt to carry out the required subtraction. As a result of the omission of this adjustment as well as that of the risk premium, our calculated rates of return are undoubtedly overevaluations.

With these observations in mind, what do our data show? To come to the central point they show that, on the average, the purchase and subsequent resale of a painting (making no allowance) for sales commissions, maintenance costs, etc.) brought an annual compounded rate of return of 0.55 percent in real terms. The median was somewhat higher: 0.85 percent. These returns are obviously far from princely. In comparison with government securities they imposed an opportunity loss upon the holder of the painting of close to two percentage points per year. That is, the rate of return on a median painting was about one-third as high as that on a government security, and the average return was only about one-sixth of the latter.

Not only were rates of return on painting as investment remarkably low, they were also remarkably dispersed, meaning that this form of investment was quite risky. Figure 1 is a histogram showing the frequency distribution of the rates of return on resales of paintings. We see that there are cases with compounded rates of return as high as 27 percent per year and others as low as -19 percent per year. In more than 40 percent of the cases returns were negative, and about 60 percent of the cases incurred an opportunity loss in the sense that they returned less than the real yield of government securities.

It may be noted that Figure 1 bears a remarkable resemblance to a normal probability distribution. This conjecture derived support from a Kolmogorov-Smirnov test of the divergence of our observed distribution from a normal distribution. Our calculation showed that the hypothesis that the two distributions are the same could not be rejected at the 0.05 percent confidence level. To that degree we can indeed conclude that art prices do behave randomly.

Figure 2 shows another attribute of our observations. The vertical axis represents length of time that elapsed between the purchase and sale of an art work, while the horizontal axis shows annual rate of return. The graph indicates that large gains or losses are experienced only by persons who hold works for a *relatively* brief period (say, less than fifty years) while as the holding period increases beyond that the range of earnings narrows and approaches very close to zero. This is, of course, what one should expect in a random process whose mean is approximately zero.

VI. More on the Possibility of Profiting Through Knowledge

It is tempting, after looking at the preceding results and the Reitlinger data, to con-

clude that investment in art is indeed perilous, but that it is dangerous primarily for the amateur who does not know what he is doing. According to this view, people who understand art, who can foresee what works will emerge triumphant from the test of time, can surely do better. Particularly the professionals who have devoted their lives to art can expect to outperform the amateur who ventures into purchasing with the temerity derived from ignorance.

Dispassionate judgement of such contentions can only give rise to skepticism. First of all, the notion that professionals are better than amateurs as prophets of price in anchorless markets is certainly belied by the well-documented performance of stock market analysts.

Beyond the caution with which the analogy with the stock market should imbue us, the evidence of the history of art connoisseurship provides strong warnings of its own. It tells us that the main lesson imparted by the test of time is the fickleness of taste whose meanderings defy prediction. Vermeer, as we know, virtually disappeared from sight for several centuries, only to be resurrected as a producer of works of the most priceless variety. El Greco is another modern rediscovery. Turner, who for a while was a leader of the British art world, is said later to have become an embarrassment to the Tate gallery because of the large collection of his works stored in their cellars; though they are now among the most valued items in the museum's collection. The pre-Raphaelites are "in" once more. Reitlinger's list of painters contains many unrecognizable names such as Wouverman, Berchem, and Van Ostade, who once were anxiously sought after but who were all but forgotten when Reitlinger wrote. Apparently some of them have again become more fashionable. Who knows if that will happen to others and, if so, when that will occur?

It is true, of course, that the profitable investments in our sample were made by those who purchased Vermeers, Turners, and pre-Raphaelites when they were not à la mode, and the heavy losers were the early buyers of Berchem, Van Ostade, and their ilk. But that is only to say that a winner is a winner and a loser is a loser. It is, perhaps, a helpful observation to the art historian, whose very legitimate metier is an exercise in hindsight. It is, however, no help to those who would foresee the future in making art purchases for investment. Only those critics who have succeeded as instruments for the redirection of general tastes seem really to have been in a position to profit from their judgement.

VII. Concluding Comment

I have argued here that if prediction as applied to stock prices is a losing game, it is certainly unlikely to be a winner in the market for works of art. Of course, none of this implies that people should desist from the ownership of art works. It may well represent a very rational choice for those who derive a high rate of return in the form of aesthetic pleasure. They should not, however, let themselves be lured into the purchase of art by the illusion that they can beat the game financially and select with any degree of reliability the combination of purchase dates and art works that will produce a rate of return exceeding the opportunity cost of their investment.

REFERENCES

Homer, Sidney, *A History of Interest Rates*, 2nd ed., New Brunswick: Rutgers University Press, 1977.

Malkiel, Burton G., *A Random Walk Down Wall Street*, New York: W. W. Norton, 1973.

Phelps-Brown, E. H. and Hopkins, Sheila, "Seven Centuries of the Prices of Consumables," *Economica*, November 1956, *23*, 296–314.

Reitlinger, Gerald, *The Economics of Taste: The Rise and Fall of the Picture Market, 1760–1960*, New York: Holt, Reinhart and Winston, 1961.

International Monetary Fund, *1949–1978 International Statistics Yearbook*, Washington, D.C., 1979.

Dance in New York: Market and Subsidy Changes

By DICK NETZER*

Let us suppose that, ten years ago, one had been possessed of perfect knowledge of the prospective state of the economy over the coming decade, and of government and foundation subsidy policy for the performing arts over that period. What would have been the corollary forecast for the state of dance in New York in 1985?

It surely would not have been optimistic. National Endowment for the Arts (NEA) Dance Program grants (for all recipients, in New York and elsewhere) were roughly constant in real terms until 1980, but subsequently have fallen sharply: the fiscal 1985 grant total was 22 percent below the fiscal 1975 level. And there was no reason to expect the NEA to concentrate its reduced resources on New York companies. Moreover, the discontinuance of the NEA Dance Touring Program—in 1975, touring had been the main economic base of all New York-based companies other than the New York City Ballet—should have endangered at least some companies. Meanwhile, the New York State Council on the Arts (NYSCA) had fared badly in its total appropriations in the wake of the fiscal problems of New York City and state: in fiscal 1984, NYSCA Dance Program allocations were 24 percent smaller in current dollars than they had been in fiscal 1975, and nearly 60 percent smaller in real terms. By 1975, the major foundations, which had had a crucial role in the shaping of dance and dance audiences in the United States, had announced their intention of substantially reducing their presence in the arts generally, and did so.

*Urban Research Center, New York University, New York, NY 10003. Frances Abbadessa and Maggi Hayes helped complete the jigsaw puzzle that data on dance comprise, and also participated in the interviewing that provided background for the observations in this paper. Among the interviewees, Cora Cahan, Eliot Feld, Mary Hays, David White, and Robert Yesselman were especially helpful.

The high rate of inflation between 1975 and 1981 would have been worrisome to our forecaster. During the previous decade, compensation levels in the performing arts had risen considerably in real terms, and much more than wage rates in the economy at large. So, very large cost increases might have been expected. After 1978, another source of cost pressure emerged: the impact of deregulation of air fares on touring costs. While deregulation resulted in considerably lower fares on some routes, mainly between pairs of large cities, there have been sharply higher fares on lower-density routes. The latter account for much of the touring travel by all but the largest companies.

Some performing arts companies regularly sell out nearly all of their performances; this is especially true of music at major festivals. Few if any dance companies are able to do this. Thus, the otherwise adverse factors might have been offset, if companies were able to sell more of those zero-marginal-cost empty seats. That would be conceivable if real incomes were rising rapidly and the income elasticity of demand for dance were high. But, as James Gapinski's paper at this session shows once more, it is not reasonable to expect high orders of income elasticity of demand for the performing arts, except possibly for modern dance companies. In any case, real incomes were *not* rising rapidly during the decade. If dance company managers had had foreknowledge of the Gapinski results for London for the 1971–83 period, there would have been considerable hesitation, at least for the smaller companies, about increasing the real price of tickets as a means of coping, for both the own-price and substitute-price elasticities are quite high for those companies.

A final unpleasant fact for New York companies was the indication in the mid-1970's that the local market for dance might be satiated by the recent increase in the number of dance performances in New

York: there were a good many empty seats. (I alluded to this at several points in the last chapter of my 1978 book.) So, in all, our forecaster would have seen a rather bleak future for dance in New York in 1985. In the event, he or she would have been dead wrong. Dance in New York in 1985 is alive and well, thriving in output terms and getting by—marginally and sometimes by what seems high-wire acts—financially. The balance of this paper describes some of the economic attributes of dance in New York now and attempts to explain the relative prosperity.

I

Economic data on dance always have been even less adequate than those for other performing arts disciplines, poor as the latter are. Since 1982, Dance/USA, a Washington-based service organization, has surveyed the larger companies, so there is fairly good data for 1983 and 1984 for these companies, but time-series data for as little as a decade are sorely lacking. The Ford Foundation study of the finances of the performing arts from 1965 through 1971, and its unpublished follow-up study through 1974 (both are analyzed in ch. 5 of my book), covered only a few dance companies, and the membership of the panel was not constant; an NEA-sponsored study done by Informatics extended the data for a constant panel from the Ford second group through 1979, and that series can be resumed for 1983 and 1984 with Dance/USA data. But the companies not covered include a number of major New York ones, making this a poor source for dealing with the New York scene.

It is possible to make a reasonable estimate of the size of the nonprofit professional dance "industry" as it now stands (or did, in 1984), and the New York share of the nationwide totals. The 31 companies covered by the Dance/USA survey for 1984, plus the Alvin Ailey company (which was not included in the survey but is the fifth largest company as measured by total operating revenue), had total operating revenues of about $95 million in 1984. There were a few other largish groups not in the survey (for example, Dance Theater of Harlem and Martha Graham), but most are quite small. Using data on number of performances, NEA grants and a number of assumptions about the other variables, the combined operating revenues of the professional dance companies not in the survey can be estimated at between $7.5 and $9.5 million in 1984 (the range indicates considerable uncertainty about earned revenue per performance) for a total of $103–105 million, of which 52–53 percent is accounted for by companies resident in New York City. In addition, there are the operations of service organizations, individuals and groups involved in the dance "industry" whose finances are not directly reflected in the books of the performing organizations as such; without double-counting receipts of presenting organizations that show up as earned revenue of dance companies, the total revenue of this sector can be estimated at $5–6 million in 1984, for a total industry size of roughly $110 million. The industry's main output is live performances: the best source lists about 4,800 of these in 1984–85, 1,500 of which were in New York City.

Table 1 contains financial data comparing the larger companies resident in New York with their counterparts elsewhere, with the New York City Ballet and the American Ballet Theater (ABT), the "giants" of the industry, shown separately. The revenue structures of the three groups differ in a few respects. ABT and the New York City Ballet are more earned-income-dependent than the other groups, by a significant margin, and the sources of earned income also differ. The non-New York companies earned more than 80 percent of their performance income from home stands (including Nutcracker performances, a substantial component of their activities), which is a great deal more than is true of New York companies. The New York companies aside from the NYC Ballet are great tourers, in the United States and abroad. As for "contributed" income, the "other" New York companies do very well with both foundations and the NEA: these companies accounted for less than 20 percent of the total attendance of all 32 companies, but nearly one-half of NEA grants and nearly one-fourth of foundation grants.

TABLE 1—PERCENT DISTRIBUTION OF REVENUE OF 32 MAJOR DANCE COMPANIES BY SOURCE, 1984[a]

	11 Companies, New York City		21 Companies, Other Cities
	NYC Ballet & ABT[b]	Other	
Total Revenue:			
Performance	62	51	55
Other Earned	5	8	7
Total Earned	67	59	62
Contributed	33	41	38
Total	100	100	100
Performance Income:[c]			
Home season	45	47	44
U.S. Tour	39	35	16
Foreign Tour	7	18	3
Nutcracker	9	0	37
Total	100	100	100
Contributed Income:			
Individuals	27	11	24
Foundations	13	27	21
Corporations	16	10	17
Other Private[d]	24	15	17
Federal Government	6	25	9
State and Local Government	14	12	12
Total	100	100	100
Exhibit: Contributions by Board, Percent of			
Total Contributed Income[c]	9	7	34
Private Contributions	11	11	43
Exhibit: Total Revenue in $ Million	$32.7	$18.2	$44.1

[a] Includes 31 companies covered by the Dance/USA annual statistical survey, which is the source of these data, plus data for the Alvin Ailey company, from its 1984 financial statement.

[b] New York City Ballet and the American Ballet Theater, shown separately because they are so much larger than other companies.

[c] Data for "other New York City" companies exclude Alvin Ailey.

[d] Mainly fund-raising events.

The non-New York companies depend very heavily on their own boards for contributions, which is not true of the New York companies.

An important aspect of the role of New York in dance is as a provider of audiences for performances by dance groups from other places, presumably audiences that are knowledgeable, critical and receptive to novel productions, simply because of the opportunities open to New Yorkers interested in dance.

TABLE 2—PROFESSIONAL DANCE PERFORMANCES IN THE UNITED STATES, JULY 1984–JUNE 1985, BY LOCATION OF PERFORMANCE AND RESIDENCE OF COMPANY

Company Residence	Location of Performance: New York City	Location of Performance: Elsewhere	Total
I. 32 Major Companies (those included in the Dance/USA 1984 survey plus the Alvin Ailey company)			
New York City	405	466	871
Elsewhere	39	899	938
Total	444	1,365	1,809
II. All Other Companies			
New York City	184	136	320
Elsewhere	883	1,797	2,690
Total	1,067	1,933	3,010

Source: Tabulated from monthly issues of *Dance* magazine.

There are indeed a very large number of performances and seats available to New Yorkers during a season. The New York companies covered by Table 1 offered just over one million seats and drew a total attendance of 842,000 in their home stands during the 1983–84 season. In addition, there were visiting performances by the major non-New York companies covered by the Dance/USA survey, home performances by New York companies not included in Table 1 (mostly but not entirely small companies), and visiting performances by dozens of dance companies not in the survey. In 1984–85, according to a tabulation of the data in the monthly directory of performances in *Dance* magazine, there were more than 1,000 such performances in New York, in addition to the 444 performances by the resident companies included in Table 1 (see Table 2). If the 1983–84 number of performances was anything like the 1984–85 schedule, these performances offered about 450,000 seats, for a total of 1.46 million.

On an average night during the main season, from late September to mid-June, there were more than 6,100 seats available for professional dance performances and, most likely, more than 75 percent of those seats were sold (many of the lesser-known companies sell far less than the 80 percent sold by the majors, but those that do so perform in the smaller houses). This suggests a very

substantial regular audience for dance. There is little useful audience survey data; opinion of insiders is that dance, like opera, is marked by habitues who attend a large number of performances in the course of a year, in contrast to the Broadway theater, with its huge audience of occasional attenders. If the mean number of performances attended is ten a year, then (at the 75 percent of capacity sold figure) the audience consists of roughly 100,000 New Yorkers.

The data in Table 2, together with the some of the financial data, suggest that dance remains very concentrated in New York. In performance terms, New York is the base for large companies that in 1984–85 offered more than one-third of all large-company performances in cities outside New York; it provides an audience for a modest number of performances by large companies resident elsewhere (only a few of these tour to New York at all), but it provides the audience for about one-third of all performances by small companies resident elsewhere; and, in sum, about one-third of all performances by all companies, large and small, were in New York last year.

One change over the past decade that contributes to this role is the increase in and improvement of the physical facilities in which dance is presented. Since the early 1970's, one of the major dance theaters (the City Center) has been much improved, a new dance theater created (the Joyce), and substantial blocks of time in two large theaters (the Opera House at the Brooklyn Academy of Music and the Metropolitan Opera House) dedicated to dance, to add to the State Theater at Lincoln Center. In addition, a number of smaller dance spaces have been developed and Broadway houses continue to be used on occasion. Obviously, these houses would not continue to be available if they were not used for dance, but there does seem to be some degree of supply creating demand, which of course is feasible for an industry that relies so much on subsidy.

II

How then can we explain the disparity between the evident prosperity and the predicted outcome? Clearly, we have misspecified the predictor variables: both market forces and subsidy policies need to be described more subtly.

In 1975, New York companies other than the NYC Ballet had very brief home seasons in New York. That was not a consequence of artistic preference or of the absence of suitable theaters (in a pinch, there were at least a dozen dark Broadway theaters available then). Instead, it was a consequence of the absence of NEA subsidy—the subsidy then was for touring, not home stands—and the high costs of the theaters, which meant that each week in New York lost money, even if high percentages of capacity were sold. Most of the major New York companies now have home seasons that are at least twice as long as they were a decade ago. They sell 80 percent or more of capacity, at significantly higher relative prices. The fragmentary national time-series data for small panels of companies suggest that ticket prices probably increased by at least one-third more than the price level between 1975 and 1984; there is no reason to think that the increase in New York was any lower.

So the audience *has* expanded, with some combination of more frequent attendance by devotees (the insiders' favorite explanation) and the creation of new devotees. Price elasticity must have been very low for the committed fans, even lower than was the case in London (there probably was no problem with regard to the substitute lively arts in New York: it appears that all their prices rose roughly in proportion). Conceivably, the audience may have been expanded because the theaters are better and more accessible.

However, it seems plausible that the audience expansion is tied to the maturation of the industry in New York, with dance now as accepted and respectable as theater, music, and opera as part of the New York scene. That respectability has brought with it a huge increase in contributions by individuals (and not board members, as in the case of the non-New York companies) to the companies with good financial records (and probably to the others as well) and also substantial giving, which is new, by corporations. The increased private giving has offset the

decline in NYSCA support, and helped pay for the deficits on the home seasons. The process is circular: the longer home stands create a physical presence that very much supports efforts to increase private giving. It is, of course, very much open to question how long the rate of increase in private giving can be maintained, with or without federal tax reform.

Another element is the change in NEA subsidy policy, with direct support of dance companies instead of indirect support via grants to presenters to subsidize touring performances. Some New York companies have been hurt, not helped, by the change, especially the very small and very fragile groups, but the dozen largest seem better off for the change. Moreover, most of them have been successful in securing NEA Challenge Grants; in those cases, NEA funding has increased, not declined, in real terms.

In a real sense, NEA and the foundations have achieved much of what they set out to do in the 1960's with respect to dance companies in New York (and this is also true of the larger ballet companies in other big cities; the demise of the Dance Touring Program may have produced some retrogression with respect to the other main goal, bringing dance to middle America): there are *institutions*, not just makeshifts centered on the choreographer, organizations that have some degree of stability, are sufficiently respectable to attract heavy private giving, continue to be creative and continue to play to national and international audiences, despite the longer home seasons.

This is not to say that all is well with dance in New York. Institutions—with their obligations to staff and for the real estate they occupy—can be even more vulnerable to economic adversity than the fringe performing arts group with no money but no obligations. Also, large size and permanence in dance means large space requirements, in the face of rapidly rising rents in New York's booming central business district. The institutions may do well, but will institutionalization, heavy private and public subsidy to the institutions, and long reasons by them, make for a hostile environment for the new and small companies which have been a principal instrument of the development of dance in the United States, and especially in New York? Does success for Eliot Feld impede the emergence of new Feld Ballet companies, by preempting audiences and donors and the attention of critics? This issue exists throughout the arts, but may be most pointed in dance, in view of the recency of its maturity.

REFERENCES

Gapinski, James H., "The Lively Arts as Substitutes for the Lively Arts," *American Economic Review Proceedings*, May 1986, *76*, 20–25.

Netzer, Dick, *The Subsidized Muse*, New York: Cambridge University Press, 1978.

Dance/USA, *Update*, August/September 1985.

Ford Foundation, *The Finances of the Performing Arts*, Vol. I, New York: Ford Foundation, 1974.

Informatics, Inc., "Growth of the Arts and Cultural Organizations in the Decade of the 1970s," Final Report to the National Endowment for the Arts, Rockville: Informatics, Inc., 1983.

The Lively Arts as Substitutes for the Lively Arts

By JAMES H. GAPINSKI*

The notion that the price of substitutes serves as a determinant of lively arts demand is hardly new. It dates back decades at least to the seminal work by William Baumol and William Bowen (1966, p. 244), who contended that movies substitute for live performances. Susan Touchstone (1980, p. 36), examining lively arts demand in the United States, followed the lead of Baumol and Bowen by defining substitute price in terms of movie admission, while Glenn Withers (1980, p. 739), also for the United States, cast it in terms of reading or recreation. I (1984, p. 462) considered both types of measures in connection with a study of the demand for Shakespeare in Great Britain. But surely if movies or reading or recreation are substitutes for the lively arts, then so are those arts themselves. Should *Richard II* become dearer, an individual might elect to attend *La Boheme* or *Fidelio* or *Swan Lake* rather than to sit through "Superman II." The lively arts are not homogeneous. Each has its own set of characteristics, and consequently substitutes lie *within* the arts spectrum. Apart from fleeting acknowledgment by, say, Alan Peacock (1981, p. 3), this point has been ignored by demand analyses to date. It is not ignored here.

I. Modeling

Denoted Q_{cr}, the quantity of cultural experiences demanded by resident r from company c satisfies the linear function

$$(1) \qquad Q_{cr} = \alpha_c + \alpha_p P_c + \alpha_S S_c + \alpha_Y Y_r,$$

where Y_r symbolizes r's real income, P_c designates the real price of an experience from c, and S_c designates the real price of a cultural experience that substitutes for one from c. This substitute is created in some other art form. Similarly, tourist t has the demand function

$$(2) \qquad Q_{ct} = \beta_c + \beta_P P_c + \beta_S S_c + \beta_Y Y_t,$$

where Q_{ct} represents the quantity of experiences demanded by t from c and where Y_t notates t's real income.

Total demand from R residents and T tourists for c's cultural experiences must be $Q_c = \Sigma_{r=1}^{r=R} Q_{cr} + \Sigma_{t=1}^{t=T} Q_{ct}$, which, after an exercise that involves inter alia restating the company-specific parameters α_c and β_c by company-specific shift variables, leaves

$$(3) \quad q_c = \alpha_1 + \sum_{j=2}^{j=C} a_j D_j + \beta_1 M + \sum_{j=2}^{j=C} b_j D_j M + \alpha_P P_c + \alpha_S S_c + \alpha_Y y_R + \beta_P P_c M + \beta_S S_c M + \beta_Y y_T M,$$

where q_c and M stand for experiences per resident Q_c/R and tourists per resident T/R, respectively, whereas y_R and y_T stand for real resident income per resident $\Sigma_{r=1}^{r=R} Y_r/R$ and the tourist counterpart $\Sigma_{t=1}^{t=T} Y_t/T$, respectively. C signifies the number of companies, and shift variables D_j are dichotomous; that is, $D_j = 1$ when $c = j$ but $D_j = 0$ otherwise.

Equation (3) constitutes a general model that can be used to produce other specifications. In particular, systematically imposing restrictions on the tourist coefficients yields five corollary equations including

$$(4) \quad q_c = \alpha_1 + \sum_j a_j D_j + \alpha_P (1+M) P_c + \alpha_S (1+M) S_c + \zeta_Y y_R,$$

*Florida State University, Tallahassee, FL 32306. Most data for this project were collected during my second tour of duty at the University's London Study Center. Thanks are owed to many for their assistance in that collection effort, but an extra tip of the bowler must go to George Darroch of the Arts Council of Great Britain. Without George's help, little would have been accomplished.

for which $\zeta_Y = \alpha_Y + \xi\beta_Y$. Behind this formulation lie the presumptions that the own-price and substitute-price coefficients do not differ between tourists and residents and that real tourist income is a proportion ξ of real resident income. It is also presumed that tourists have no predetermined demand β_1 or b_j for cultural experiences.

The six equations, expression (3) and the five corollaries, never restrict the full set of tourist coefficients to zero because, as Harry Kelejian and William Lawrence (1980, pp. 337, 345) along with the Society of West End Theatre (1983, p. 4) observe, tourism is important for lively arts demand. In recognition of findings by Withers (pp. 740–41) and myself (p. 462), the six do not allow for partial adjustment. Moreover, they apparently do not require the use of simultaneous-equation estimating techniques. Withers (p. 737) argues theoretically and Thomas Moore (1968, pp. 170–72) affirms empirically that the demand for lively arts is free from simultaneity, and therefore single-equation methods are taken to be appropriate.

II. The Data and Their Processing

London provides the context for testing the six equations although, given the comparability of the London and New York cultural environments, that choice does not seem to entail much loss of generality. The period, determined by the availability of company data, covers the twelve financial years 1971–72 to 1982–83. A financial year runs from April to March.

Thirteen companies are considered. They belong to the four art forms theater, opera, symphony, and dance, and to preserve confidentiality they are referenced by code. Representing theater are *THE1* and *THE2*; opera, *OPE1* and *OPE2*. Symphony consists of *SYM1* to *SYM4* while dance has *DAN1* to *DAN5*. Information on the thirteen originates from the Arts Council of Great Britain and from the annual reports of the London Orchestral Concert Board Limited.

Twelve years of data on thirteen companies mean a sample size N of 156 observations. In that data file paid attendance, calibrated in units, quantifies cultural experiences. Nominal own price, expressed in unit pounds, is calculated by dividing attendance into box-office revenue and includes the value-added tax. Nominal substitute price, also expressed in unit pounds, is defined consonant with the proposition advanced in the introduction; namely, that arts consumers look to other art forms in contemplating substitutes. Specifically, for any company it is the unweighted average of the nominal own prices charged by the companies in the other arts. For example, *THE1*'s substitute price is the average of the prices posted by the eleven organizations comprising opera, symphony, and dance. Likewise, *OPE1*'s substitute price is the average of the prices listed by the eleven in theater, symphony, and dance.

Numbers on London's nominal disposable income, in unit pounds, are supplied by the Central Statistical Office (CSO), while those on its resident counts, in units, are extracted from *Population Trends* prepared by the Office of Population Censuses and Surveys. The city's tourist information—headcounts and nominal expenditures—is the London Tourist Board's. Headcounts are measured in units; expenditures, in unit pounds. The latter series replaces the tourist income variable, for which there are no data. The implicit assumption underlying this switch is that vacation spending is proportional to income on balance. Conversion of all nominal magnitudes into reals is done by the Retail Price Index. Based at unity for calendar year 1975, it comes from *Economic Trends* of the CSO. Conversion of a calendar-year series into a financial-year format occurs prior to any use of that series.

The joint time-series and cross-section nature of the data advises that allowances be made for both autocorrelation and heteroscedasticity. This end can be served by a slightly modified version of Jan Kmenta's (1971, pp. 509–11, 525) two-step regression procedure that enables the autocorrelation and heteroscedasticity correction factors to vary across establishments. The method treats autocorrelation first and heteroscedasticity second. High intercorrelations among the explanatory variables recommend that a third

step be taken, this one to guard against multicollinearity. Step 3 calls upon ridge regression, which adds to the cross-product matrix for the regressors an orthogonal matrix weighted by a nonnegative scalar μ.

Each of the six equations is subjected to this three-step estimation process, and the results are examined to select one of the six as the preferred formulation. The selection rule, which combines the maximum $\bar{R}^2$ criterion (equivalently, the minimum residual-variance norm) with a requirement that the equation have correct signs for the coefficients of the usual demand determinants (own price and income), eventually points to equation (4).

III. The Findings

Table 1 reports that the fit of equation (4) is quite respectable in terms of the overall measures and that thirteen of the sixteen coefficients are significant. More precisely, the own-price coefficient α_P is significant, its Student t-value being -2.594. The income coefficient ζ_Y is not, however. This insignificance could not be reversed by ridge regression, and hence its explanation may rest more in theory than in data: at least high intercorrelations seem to be blameless.

In confirmation of the fundamental hypothesis that the lively arts substitute for the lively arts, price coefficient α_S proves to be positive and significant with a Student t-value of 2.209. Each art form competes against the others, and companies find themselves to be price interdependent across forms.

Shift coefficients a_j for *THE*2 and for both opera concerns are positive whereas those for the organizations in symphony and dance are consistently negative: *ceteris paribus*, arts consumers frequent theater or opera more than symphony or dance. The least frequented enterprises happen to be *DAN4* and *DAN5*, a finding compatible with the case made by Baumol and Bowen (pp. 253–57) that attendance plummets when contemporary works are performed. *DAN4* and *DAN5* are unquestionably the most heavily contemporary of the dance troupes and almost unquestionably the most heavily contemporary of all thirteen.

TABLE 1—EMPIRICAL RESULTS FROM EQUATION (4)

		Elasticity with Respect to Real		
Reference	Parameter Estimate[a]	Own Price	Sub. Price	Income per Resident
THE1	.0329	−.10	.18	.10
THE2	.0385	−.05	.09	.05
OPE1	.0079	−.12	.13	.08
OPE2	.0074[b]	−.25	.15	.09
SYM1	−.0249	−.19	.44	.22
SYM2	−.0265	−.32	.55	.28
SYM3	−.0289	−.35	.65	.33
SYM4	−.0274	−.24	.53	.27
DAN1	−.0143	−.18	.28	.15
DAN2	−.0035[b]	−.21	.21	.11
DAN3	−.0331	−.64	1.10	.58
DAN4	−.0381	−.81	2.28	1.21
DAN5	−.0381	−.70	2.06	1.09
α_P	−.57-3			
α_S	.85-3			
ζ_Y	.23-5[c]			
Theater		−.07	.12	.06
Opera		−.18	.14	.09
Symphony		−.27	.53	.27
Dance		−.29	.50	.26

Notes: Overall regression measures, based on $N=156$, read $R^2=.95$, $\bar{R}^2=.94$, and $F_{15,140}=164.60$ with $\mu=0$. Parameter estimates listed by company code are the a_j excepting that for *THE1*. It is α_1.

[a] $\pm A-B=\pm A\cdot 10^{-B}$; for example, −.57-3 becomes $-.57\cdot 10^{-3}$.

[b] Insignificant on a 5 percent two-tail Student t-test.

[c] Insignificant on a 5 percent one-tail Student t-test.

The elasticities posted in Table 1 hold at the means for the pertinent variables. Own-price elasticity satisfies $\alpha_P(1+M)P_c/q_c$ while substitute-price elasticity and income elasticity satisfy $\alpha_S(1+M)S_c/q_c$ and $\zeta_Y y_R/q_c$, respectively.

Regarding own price, demand is uniformly inelastic. It is most inelastic for *THE2* and *THE1*, and least so for *DAN3*, *DAN5*, and *DAN4*. Substitute-price elasticity follows a similar sequence although it includes for the aforementioned dance trio numbers in excess of unity. Those companies are particularly sensitive to pricing actions in the other arts. When viewed collectively the company figures indicate that theater and opera have the smaller own-price and substitute-price elasticities, while symphony and

dance have the larger. On balance, theater and opera feel less audience response to price changes initiated by themselves or by their competitors; conversely, symphony and dance feel more.

The income elasticities, which perhaps should be regarded with extra caution due to the insignificance of ζ_Y, tend to remain below unity and run counter to the impression that a cultural experience is a luxury. Yet they agree nicely with evidence provided by Steven Globerman (1978, p. 33), Kelejian and Lawrence (pp. 345–46), Moore (pp. 172, 175), Dick Netzer (1978, p. 29), and Withers (pp. 739–40), and they can be explained in terms of time-allocation logic. A cultural experience is a time-intensive consumable. As income rises, individuals desire more experiences, but they also value time more preciously and therefore desire less time-expensive kinds of consumption: they simultaneously seek more and fewer experiences. The net effect rules, and consequently attendance may well be income insensitive. Adding the time cost of ancillaries, including that of commutation, only strengthens the reasoning for low income elasticities. That same reasoning, of course, can explain the insignificance of ζ_Y.

What would happen to attendance at company c if a single competitor lowered its price by, say, 10 percent? Because c has numerous competitors, the substitute price bearing on it must fall by less than that percentage, and thus a straightforward application of the substitute elasticities in Table 1 to a 10 percent figure would exaggerate the outcome. Table 2 addresses this question through a matrix of changes occasioned by a 10 percent price cut initiated by one company alone. Represented there is the largest drawer for each art type (*THE2*, *OPE1*, *SYM1*, and *DAN2*) together with the smallest drawer regardless of type (*DAN4*). Each row label identifies the real price which is being decreased by 10 percent, and each column label identifies the company to which the effects apply. The top number in a cell gives the change in real substitute price facing the designated company while the bottom entry registers the change in its attendance level, all changes being reckoned from the means. For instance, a 10 percent drop in P_{THE2}, the real price charged by *THE2*, makes *OPE1* note a decline of .97 percent in its real substitute price and a reduction of 400 persons in its annual attendance.

TABLE 2—A PRICE INTERDEPENDENCY MATRIX

Impulse	Consequent Response in				
10 Percent Cut in	Real Substitute Price for Annual Attendance for				
	THE2	*OPE1*	*SYM1*	*DAN2*	*DAN4*
P_{THE2}	–	–.97	–.94	–1.14	–1.14
	281	– 40	– 48	– 57	– 56
P_{OPE1}	–1.01	–	–1.22	–1.46	–1.46
	– 49	370	– 63	– 73	– 73
P_{SYM1}	–.60	–.77	–	–.91	–.91
	– 29	– 32	223	– 45	– 45
P_{DAN2}	–1.38	–1.69	–1.64	–	–
	– 66	– 70	– 84	501	–
P_{DAN4}	–.47	–.63	–.59	–	–
	– 23	– 26	– 30	–	176

Note: Substitute-price response is expressed in percent; attendance response, in tens of persons.

Of the patterns to emerge from the table, one reveals that the price interdependencies for the large companies are strongest between *THE2*, *OPE1*, and *SYM1* on the one hand, and *DAN2* on the other. A price rollback by any member of the Big Three causes *DAN2* to suffer the greatest attendance loss, and correspondingly the greatest attendance loss for any of the Three comes at the hands of *DAN2*. Another pattern discloses that each Big Three organization exerts more influence on little *DAN4* than *DAN4* exerts in return. This property has special intuitive appeal.

Possibly the clearest pattern to emerge involves the basic strength of the substitution phenomenon in this one-on-one situation: a price change by a single company alone has minor impact on a second company. The action of one concern is swamped by the inaction of all others to the point where a 10 percent cut in own price never drops substitute price by more than 1.69 percent. It follows that the greatest attendance response to a price maneuver occurs for the initiating firm itself. A 10 percent reduction in P_{THE2} boosts *THE2* attendance by 2,810 patrons but erodes a competitor's by at most 570.

Obversely, a 10 percent slash in P_{OPE1}, P_{SYM1}, P_{DAN2}, or P_{DAN4} has a much smaller consequence for *THE2* attendance than does the adjustment in P_{THE2}. A company's own pricing policy holds substantially more sway over its opportunities than does the pricing policy of an individual competitor.

Circumstances change drastically when rivals act together. When all companies in the other arts revise their prices downward by 10 percent, *THE2* loses 4,800 patrons annually. That loss is almost twice what *THE2* can prompt on its own. Under like conditions, *OPE1* performs before 4,160 fewer enthusiasts each year, and *SYM1*, *DAN2*, and *DAN4*, respectively, play to 5,150, 5,020, and 4,980 fewer. *DAN4*, shown in Table 1 to be especially vulnerable to competition, has a loss factor of almost three. Price interdependencies do matter.

IV. Concluding Remarks

Attendance and the associated box-office revenue are essential to any arts company, and the findings presented here relate to several possibilities for increasing them. For instance, the price inelasticity of demand, characteristic of the thirteen companies without exception, implies that revenue can be increased by increasing price. Attendance necessarily declines in the process. But because companies are price interdependent, one company's attendance loss is partly another's gain, and with that gain comes increased revenue: One company's price increase raises not only its own revenue, but also the revenues of others elsewhere in the cultural community. An obvious prospect, then, is a price increase.

Another derives from the unresponsiveness of attendance to income. This insensitivity reflects the arts consumer's sensitivity to the use of time and suggests that reducing time cost would push attendance and revenue upward. Proposals to eliminate Acts IV and V of *Julius Caesar*, to restrict intervals to five minutes, or to abolish overtures and encores can be summarily dismissed. Not easily dismissed are recommendations for British Rail and London Transport to increase the frequency of trains into, within, and from the West End around performance times to reduce the delays at boarding and transfer points. Nor easily dismissed are recommendations for increasing the number of car parks in the West End, or for disentangling West End traffic flows by redesignating principal arteries as one-way thoroughfares.

Such proposals hold promise for improving the economic well-being of arts companies and for diminishing the reliance on gifts, either private or public. However, they may not be very prudent or practical when judged on a larger scale. Increasing train service would necessitate more public expenditures as would the building of car parks and the rerouting of streets. Would these costs be outweighed by the benefits? Costs aside, the challenge of persuading British Rail or London Transport to alter schedules may be destined to fail from the start. Thus, although lively arts companies have the potential for improving their economic lot, they may be able to do so only at the margin.

REFERENCES

Baumol, William J. and Bowen, William G., *Performing Arts—The Economic Dilemma*, New York: Twentieth Century Fund, 1966.

Gapinski, James H., "The Economics of Performing Shakespeare," *American Economic Review*, June 1984, *74*, 458–66.

Globerman, Steven, "Price Awareness in the Performing Arts," *Journal of Cultural Economics*, December 1978, *2*, 27–41.

Kelejian, Harry H. and Lawrence, William J., "Estimating the Demand for Broadway Theater: A Preliminary Inquiry," in William S. Hendon et al., eds., *Economic Policy for the Arts*, Cambridge: Abt Books, 1980, 333–46.

Kmenta, Jan, *Elements of Econometrics*, New York: Macmillan, 1971.

Moore, Thomas Gale, *The Economics of the American Theater*, Durham: Duke University Press, 1968.

Netzer, Dick, *The Subsidized Muse*, London: Cambridge University Press, 1978.

Peacock, Alan, "Economics, Inflation and the Performing Arts," mimeo., University Col-

lege at Buckingham, England, August 1981.

Touchstone, Susan Kathleen, "The Effects of Contributions on Price and Attendance in the Lively Arts," *Journal of Cultural Economics*, June 1980, *4*, 33–46.

Withers, Glenn A., "Unbalanced Growth and the Demand for Performing Arts: An Econometric Analysis," *Southern Economic Journal*, January 1980, *46*, 735–42.

Central Statistical Office, *Economic Trends*, Annual Supplement, London: HMSO, December 1983.

London Orchestral Concert Board Limited, *Annual Report of the Board of Management*, London, various years.

Office of Population Censuses and Surveys, *Population Trends*, London: HMSO, Winter 1983.

Society of West End Theatre, "West End Theatre Data—1982," mimeo., London, March 1983.

Supply Side Economics: Old Truths and New Claims

By MARTIN FELDSTEIN*

Experience has shown that the notion "supply-side economics" is a malleable one, easily misused by its supporters, maligned by its opponents, and misinterpreted by the public at large. Perhaps now, five years after supply-side economics became a slogan for a changing economic policy, it is possible to assess what supply-side policy really means and how the policies adopted under that banner have fared.

The term supply-side economics originated as a way of describing an alternative to the demand side emphasis of Keynesian economics. The essence of Keynesian analysis is its conclusion that the level of national income and employment depend on the level of aggregate demand, and that easy money and expanded budget deficits, by stimulating demand, can increase output and employment. Although this may have been an appropriate emphasis during the depression years of the 1930's when Keynes developed his theory, by the 1960's and 1970's it was clear to most economists that it was wrong to focus exclusively on demand and to ignore the factors that increase the potential supply of output—capital accumulation, technical progress, improvements in the quality of the labor force, freedom from regulatory interference, and increases in personal incentives. Many of us also concluded that the persistently high level of measured unemployment did not reflect inadequate demand but was due to government policies like unemployment insurance, welfare restrictions, and the minimum wage that reduced the effective supply of labor.

In all of these ways, many of us were supply siders before we ever heard the term supply-side economics. Indeed, much of our supply-side economics was a return to basic ideas about creating capacity and removing government impediments to individual initiative that were central in Adam Smith's *Wealth of Nations* and in the writings of the classical economists of the nineteenth century. The experience of the 1930's had temporarily made it easy to forget the importance of the supply factors, but by the 1970's they were returning to the mainstream of economics. (See my 1981, 1982 papers.)

It is important in any discussion of supply-side economics to distinguish the traditional supply-side emphasis that characterized most economic policy analysis during the past 200 years from the new supply-side rhetoric that came to the fore as the decade began.

I. The Shift in Policy

Economic policy took a few hesitating steps in the traditional supply-side direction in the late 1970's with deregulation in the transportation industry, a significant reduction in the tax on capital gains, and the partial taxation of unemployment compensation. But it was only in 1981 that Congress enacted the major tax bill that has become the centerpiece of supply-side economics.

The emphasis throughout that tax legislation was on changing marginal tax rates to strengthen incentives for work, saving, investment and risk taking. For individual taxpayers, the basic features of the Economic Recovery Tax Act of 1981 were a 25 percent across-the-board reduction in personal tax rates, an extra tax reduction for two-earner

†*Discussants*: Barry P. Bosworth, The Brookings Institution; Manuel H. Johnson, U.S. Department of the Treasury; Victor A. Canto, University of Southern California.

*Professor of Economics, Harvard University, Cambridge, MA 02138, and President, National Bureau of Economic Research.

families, an increased exemption for long-term capital gains, and the creation of universal Individual Retirement Accounts that effectively permit the majority of American employees to save as much as they want out of pretax income and pay tax on those savings on a consumption tax basis. Personal tax brackets were also indexed to prevent inflation from raising real tax burdens (although this indexing was only scheduled to begin in 1985). For businesses, the 1981 legislation contained accelerated depreciation schedules that significantly reduced the cost of investment in plant and equipment, and an increased tax credit for research and development.

The Reagan Administration also began an unprecedented reversal of the share of *GNP* absorbed by government nondefense spending. Those outlays declined from 15.1 percent of *GNP* in fiscal year 1980 to 14.1 percent of *GNP* in FY 1984. When the Social Security and Medicare outlays are excluded, this spending declined from 9.3 percent of *GNP* in 1980 to 7.4 percent in 1984. These spending reductions were significant not only because they released resources that could be used to finance tax rate reductions, but also because they were often achieved by shrinking programs that in themselves had adverse incentive effects.

President Reagan also provided strong support for the anti-inflationary Federal Reserve policies. The sharp fall in inflation between 1980 and 1982 significantly reduced the effective tax rates on the return to corporate capital, increasing the real after-tax return to savers as well as reducing the uncertainty of saving and investment.[1]

II. Excessive Claims

These policies were a major step in the direction recommended by supply-side economists of both the new and old varieties. What distinguished the new supply siders from the traditional supply siders as the 1980's began was not the policies they advocated, but the claims that they made for those policies.

The traditional supply siders (although I dislike labels, I consider myself one of that group) were content to claim that the pursuit of such tax, spending, and monetary policies would, over the long run, lead to increased real incomes and a higher standard of living. We recognized that the key to this process was increased saving and investment and knew that that would take a long time to have a noticeable effect.[2]

The "new" supply siders were much more extravagant in their claims. They projected rapid growth, dramatic increases in tax revenue, a sharp rise in saving, and a relatively painless reduction in inflation. The height of supply-side hyperbole was the "Laffer curve" proposition that the tax cut would actually increase tax revenue because it would unleash an enormously depressed supply of effort. Another remarkable proposition was the claim that even if the tax cuts did lead to an increased budget deficit, that would not reduce the funds available for investment in plant and equipment because tax changes would raise the saving rate by enough to finance the increased deficit. It was also claimed that the rapid rise in real output that would result from the increased incentive to work would slow the rate of inflation without the need for a rise in unemployment because the increased supply of goods and services could absorb the rising nominal demand.

Probably no single individual made all of those claims—at least not at the same time. And anyone who feels the need to defend his name can argue that the administrations's 1981 economic program was not enacted exactly as proposed. Nevertheless, I have no doubt that the loose talk of the supply-side

[1] The effects of inflation on effective tax rates on investment in plant and equipment are analyzed in the papers collected in my book (1983a).

[2] Some of us were also nervous about the magnitude of the enlarged tax cut that emerged from the bargaining between the congressional Democrats and Republicans. I advocated making a large part of the personal tax cut an immediate indexing of the tax brackets (to eliminate the risk of a real tax cut that was either bigger or smaller than needed to offset bracket creep during the years 1981–85) and phasing in much of the remaining tax cut only as spending cuts were achieved.

extremists gave fundamentally good policies a bad name and led to quantitative mistakes that not only contributed to subsequent budget deficits, but also made it more difficult to modify policy when those deficits became apparent.

III. Growth and Recovery

To assess the claims of the new supply siders, it is useful to compare the actual growth of real *GNP* between 1981 and 1985 with the growth that the supply siders initially projected. The record shows that real *GNP* increased 10.9 percent between 1981 and 1985, only slightly more than half of the 19.1 percent predicted in the Reagan Administration's original economic plan.[3]

This 45 percent shortfall in economic growth cannot be blamed, as some of the new supply siders would now do, on a failure of the Federal Reserve to supply as much money and credit as the plan originally envisioned. The 1981 *Program for Economic Recovery* assumed that "the growth rates of money and credit are gradually reduced from the 1980 levels to one-half those levels by 1986" (p. 23) while the actual money growth rates have hardly declined at all since 1981.

Although the original forecast of nearly 5 percent a year real growth from 1981 to 1985 was improbable on the basis of both historic experience and economic theory, the shortfall was clearly exacerbated by the recession that depressed *GNP* from the third quarter of 1981 until the final quarter of 1982. The new supply siders were naively optimistic when they claimed that the double digit inflation of 1980 and 1981 could be halved in a few years without any increase in unemployment simply by increasing output enough through improved incentives to absorb the excess demand.

Most of the new supply siders have now conveniently forgotten the substantial discrepancy between their growth forecast and the subsequent experience. But some of the supply-side extremists even claim that the recovery was delayed because individuals preferred to "consume leisure" and were waiting to return to work until the final stage of the tax rate reduction had occurred. Anyone who believes that that explains the 10.7 percent unemployment in December 1982 has not studied the data on the composition and timing of unemployment or on the relation between the spending upturn and subsequent reductions in unemployment. And those who wish to believe that the cut in the tax rate stimulated a major increase in the number of people wanting to work will be disappointed by the data on labor force participation rates.

During the first four quarters of the recovery, real *GNP* increased at about the average pace of the previous recoveries. In the second year of the recovery, the rise in *GNP* exceeded the past norm. But now, eleven quarters after the recovery began, the cumulative rise in *GNP* has settled back to the middle of the range of past recoveries.

How much of the recovery has been due to the stimulus to increased supply that was provided by the new policies?[4] I have already commented on the lack of evidence of an induced increase in the number of people wanting to work. But it would be equally wrong to view the recovery as the result of the fiscal stimulus to demand as some traditional Keynesians have done (for example, James Tobin, 1984).

In fact, the rise in nominal *GNP* since 1982 can be more than fully explained by the traditional relationship to the lagged increase in money (*M*1). The division of the nominal *GNP* increase between *GNP* and inflation was, however, more favorable than would have been expected on the basis of past experience; somewhere around 2 percent of the 15 percent rise in real *GNP*, since the recovery began cannot be explained by the increase of nominal *GNP* and the past pattern of inflation and might therefore be attributed to supply side factors. However, the rise in the exchange rate fully explains

[3] See The White House, page S-1. This official forecast predicted less growth than some of the more ardent new supply siders anticipated.

[4] The remainder of this section is based on my 1986 article.

the relatively favorable inflation experience and leaves no unexplained rise in real *GNP*. Of course, it might be argued that supply-side factors contributed to the dollar's rise. Only further research will resolve whether supply side influences have contributed to the rise in real *GNP* since 1981.

Let me emphasize that, to a traditional supply sider like me, the positive but apparently modest supply-side effect is neither surprising nor disappointing. Although we would expect some increase in work effort from the reduction in the highest marginal tax rates, past evidence all points to relatively small changes. The favorable effects of improved incentives for saving and investment can only be expected after a much longer period of time.

IV. Tax Revenue

Perhaps the most dramatic claim of some of the new supply siders was that an across-the-board reduction in tax rates would be self-financing within a few years because of the increased output that results from the enhanced after-tax pay.[5] It is, of course, very difficult to disentangle the effects of the tax legislation from other things that influenced tax revenue. But a very careful study by Lawrence Lindsey (1985a, b) indicates that in 1982 the response of taxpayers did offset about one-third of the effect of the tax cut on federal receipts.

Lindsey reports that about 65 percent of the induced offsetting rise in tax revenue reflects higher pretax wages, salaries, and business profits than would have been anticipated without the change in tax rates and tax rules, 25 percent reflects an increase in realized capital gains, and the remaining 10 percent is due to reductions in various itemized deductions. These induced offsetting effects are very small among taxpayers with incomes below $20,000. Only among taxpayers whose initial marginal tax rates exceeded 50 percent was there evidence that the rate reduction did not reduce federal revenue at all.

Only time will tell whether this first-year tax response overstates the long-term effect (because it reflects a shift in the timing of income receipts and deductions rather than a more fundamental change in behavior) or understates the long-term effect (because it takes time for taxpayers to adjust their behavior to new tax rules). But the effect for 1982 is clearly an economically significant one. Although the increase in taxable income fell far short of the claims made by the overoptimistic new supply siders and may have been due in large part to a restructuring of income (for example, from fringe benefits to cash) rather than an increase in work effort, the rise in taxable income is a reminder that the traditional revenue estimation method that ignores the behavioral response to tax changes can be very misleading (see my 1983b report).

V. Conclusion

The experience since 1981 has not been kind to the claims of the new supply-side extremists that an across-the-board reduction in tax rates would spur unprecedented growth, reduce inflation painlessly, increase tax revenue, and stimulate a spectacular rise in personal saving. Each of those predictions has proven to be wrong.

But it would be unfortunate if this gave a bad reputation to the traditional supply-side verities that the evolution of a nation's real income depends on its accumulation of physical and intellectual capital and on the quality and efforts of its workforce. Moreover, nothing about the experience since 1981 would cause us to doubt the time-honored conclusion of economists that tax rules influence economic behavior and that high marginal tax rates reduce incentives.

Indeed, the evidence suggests that the reduction in tax rates did have a favorable effect on work incentives and on real *GNP*, and that the resulting loss of tax revenue was significantly less than the traditional revenue estimates would imply. Traditional supply-

[5]The administration never made such a claim although the unusually strong real growth that it predicted for the first five years would have been sufficient to recoup between one-half and three-quarters of the proposed 30 percent tax cut.

side considerations are undoubtedly important in the design of economic policies in general and of tax policies in particular. But the miraculous effects anticipated by some of the new supply-side enthusiasts were, alas, without substance.

REFERENCES

Feldstein, Martin, "The Retreat from Keynesian Economics," *The Public Interest*, Summer 1981, *64*, 92–105.

______, "The Conceptual Foundations of Supply Side Economics," in *Supply Side Economics in the 1980's*, proceedings of a conference sponsored by the Federal Reserve Bank in Atlanta and the Emory University Law & Economics Center, May 1982.

______, (1983a) *Inflation, Tax Rules and Capital Formation*, Chicago: University of Chicago Press, 1983.

______, (1983b) *Behavior Simulation Methods in Tax Policy Analysis*, NBER Project Report, Chicago: University of Chicago Press, 1983.

______, "The 1983 Economic Recovery: Lessons for Monetary and Fiscal Policy," forthcoming, 1986.

Lindsey, Lawrence, (1985a) "Taxpayer Behavior and the Distribution of the 1982 Tax Cut," NBER Working Paper No. 1760, 1985.

______, (1985b) "Estimating the Revenue-maximizing Top Personal Tax Rate," NBER Working Paper No. 1761, 1985.

Tobin, James, "Unemployment in the 1980s: Macroeconomic Diagnosis and Prescription," in Andrew Pierre, ed., *Unemployment and Growth in the Western Economies*, New York: Council on Foreign Relations, 1984.

The White House, *America's New Beginning: A Program for Economic Recovery*, Washington: USGPO, February 18, 1981.

Economic Surprises and Messages of the 1980's

By LAWRENCE CHIMERINE AND RICHARD M. YOUNG*

During the first half of the decade of the 1980's, the United States and the global economy have experienced dramatic shifts in the economic environment which have been surpassed only during the Great Depression and world wars. Many, if not most, of these shifts were unanticipated by the economics profession. While it is perhaps premature to evaluate the causes of the professional misadventures that led to these "surprises," we have been asked to speculate on their nature and the accompanying "messages" that we might draw from them.

While we might have identified a variety of issues, we have chosen to focus on four not unrelated events: 1) the sharp recession and recovery of the U.S. economy; 2) the high level of real interest rates maintained throughout much of the recent past; 3) the sharp increase in the value of the dollar; and 4) the weakness in world oil prices.

We have not done a systematic survey of forecasts for these events during the late 1970's and early 1980's, but it is our impression that few analysts would have forecast any of them, although we have no doubt that, having taken that stand, many will come forward to refute us. Nonetheless, we must admit to some surprise at events and believe an initial examination of the apparent lessons we might learn from that surprise might be revealing.

I. Recession and Recovery

Perhaps the least surprising surprise has been the cyclical behavior of the U.S. economy. The scale of decline and recovery were unprecedented in postwar experience, but many might maintain that if we had been told in 1980 what fiscal and monetary policies would be pursued during 1981–85, we could have forecast the direction, if not the timing and scale, of these changes. The real surprise has not been just the economy itself, but the policies which have driven and continue to drive that economy.

In 1979 and 1980 it would have been daring indeed to forecast:

1) A monetary policy that would push some short-term interest rates to near 20 percent at their peak and would maintain short-term rates at an average some 5–8 percent above the inflation rate for more than five years, with real long-term rates that appear to have been even higher.

2) A fiscal policy that would systematically push the federal deficit to $200 billion and threaten to keep it there in the name of improving the supply responses of the economy.

Policy has surprised us, but even if the outcome of that policy might have been predicted by the more astute among us, the lessons we should take away from this experience are still debatable.

Supply siders have argued that supply-side economics, especially the tax cuts enacted in 1981, has been a roaring success. The primary evidence used to support this conclusion is the "spectacular" recovery in 1983 and 1984, and the sharp decline in inflation during the last several years. In our view, however, a careful evaluation of recent performance would lead to different conclusions, namely that the tax cuts have as yet had no material impact on the aggregate supply curve. During the 1982–84 period these tax cuts offered a clear-cut Keynesian stimulus to demand which could have been anticipated, but by now this impact has been vitiated, and by causing enormous budget deficits, excessive tax cuts have actually become counterproductive. Furthermore, the strength of the economy promoted by these tax cuts has been overstated by many supply siders and others.

*Chairman and Chief Economist, Chase Econometrics, Bala Cynwyd, PA 19004–1780, and President, Monetary Policy Forum; and Vice President and Director of Macroeconomic Services, Chase Econometrics, respectively.

The misinterpretation of economic performance has in part resulted from a failure to distinguish between the direction and the level of economic activity—while the recovery in 1983 and the first half of 1984 was strong in terms of magnitude of increase, the level of economic activity was still considerably below its potential. This reflects the extremely weak conditions from which the recovery began because of the severity of the 1981–82 recession, and the fact that it followed so closely on the heels of the previous one. In fact, unemployment, capacity utilization, profits, and other important measures of economic performance were still far from satisfactory in mid–1984, and in most cases, had not even returned to the relatively sluggish levels which existed in the late 1970's. Several industries and geographic areas were particularly depressed (and still are), having experienced virtually no recovery at all, indicating both a high degree of imbalance in addition to the far from healthy overall picture.

Moreover, the tax cuts have been given too much credit for the recovery. The dramatic easing in monetary policy that was initiated in the summer of 1982 (as indicated by both declining interest rates and rapid growth in the money supply); the favorable effect of declining oil prices, and generally low inflation, on household purchasing power; and cyclical factors, such as the huge pent-up demands which were created during the two prior recessions and the inventory cycle, were also key ingredients. Our estimates indicate that these factors contributed at least as much to the magnitude of the economic recovery as did the tax cuts (see Table 1).

TABLE 1

Factor	Estimated Contribution to Recovery
Cyclical	25–30 Percent
Lower Inflation	15 Percent
Fiscal Stimulus	30–35 Percent
Lower Interest Rates	25–30 Percent

Perhaps the key issue is that there does not appear to have been any significant improvement in long-term potential economic growth as a result of the tax cuts, contrary to predictions made by many supply siders. In particular: (a) the growth in the labor force has slowed markedly since 1980, in part reflecting a slowdown in the rate of increase in the participation rate—this has occurred despite the reduction in marginal tax rates which was supposed to stimulate more work effort; (b) the rate of increase in productivity has actually been below increases during most previous recovery periods, and has been particularly weak since mid-1984; (c) the personal saving rate during the last four years has actually been significantly below previous years, even after adjusting for demographic changes and other factors, despite sharp cuts in marginal tax rates, the enactment of various saving incentives, and extremely high real interest rates; (d) profits as a share of national income have still not returned to the peak levels of the late 1970's, and fell during 1985; (e) real net investment as a share of *GNP* is also still below previous peaks, because the strong growth in investment in 1983 and 1984 followed very depressed levels during the prior recession, and because investment in recent years has been weighted toward short-lived assets; and (f) the U.S. competitive position in world markets has been at its worst level in many decades. The faster-than-expected recovery in its early stages was thus primarily due to the reemployment of idle resources at a faster rate than had been expected, rather than to more favorable long-range growth prospects.

Finally, growth has slowed sharply since mid-1984, to about a 2.5 percent annual rate, despite the fact that the recovery was far from complete, and despite still enormous deficits. In our view, these growing deficits have been a major factor causing interest rates to remain well above historical levels, and, because high interest rates have led to an increase in net foreign demand for U.S. assets, they have also been a principal cause of the overvaluation of the U.S. dollar in recent years. Interest rates have been especially high when measured relative to the inflation rate for goods (which strongly influences capital spending and inventory decisions) and relative to wage growth (which

affects the demand for housing). The overly strong dollar exchange rate has also restrained economic activity in the United States in several ways: (*i*) It has been the major factor behind the very sharp and widespread increases in import penetration, and relatively weak exports, which have produced enormous U.S. trade deficits despite falling oil imports. (*ii*) The strong dollar has caused a profit squeeze in many industries by preventing many companies from raising prices; this in turn has reduced the growth in capital spending. (*iii*) Many companies have increased their efforts to cut wages in order to at least partially offset declining profits—this, combined with the direct job loss in the relatively high-wage manufacturing sector, has led to a sharp deceleration in the growth in personal income, and thus to slower growth in consumer spending.

Interest rates and the U.S. dollar have thus been too high to permit more rapid economic growth and are therefore the two principal factors preventing a faster completion of the recovery process—in turn, both are primarily caused by high and rising federal budget deficits at a point in the recovery when such deficits should be falling sharply. Federal deficits, and the tax cuts that have largely caused them, have thus became counterproductive for economic growth—their direct stimulus was being outweighed by the adverse effects of the excessively high interest and dollar exchange rates which they caused.

It is true that there has been a sharp deceleration in inflation during the last several years. In part, this was due to the temporary surge in economic growth, since it permitted a cyclical bulge in productivity growth which held down unit labor costs. However, our studies show that the main factors behind the slowdown in inflation are: (a) the decline in energy prices, following the explosion of such prices during the 1970's; (b) the lingering effects of extremely high unemployment; (c) the effects of the overvalued dollar and deregulation, which have made the economy much more competitive; and (d) a glut of industrial and food commodities, in part caused by the need of many *LDC*s to generate more foreign exchange. There is no evidence to support the notion that increased supplies of commodities or finished goods have resulted from lower marginal tax rates, or that the tax cuts have been a major factor in reducing inflation because of any change in the aggregate supply curve.

We can speculate about the numerous policy lessons that can be learned from the experiences of the last several years. Chief among the observations that we believe time is likely to confirm are:

Supply responses of labor to shifts in marginal tax rates are low in the short run and likely to take a minimum of 10 to 15 years to become apparent. We do not deny the theoretical plausibility of such responses, only their empirical relevance during a period when income effects are likely to dominate relative price impacts on work/leisure decisions, particularly when marginal decisions are "lumpy." And, in view of the lack of sufficient evidence to demonstrate that lower marginal tax rates stimulate saving and investment, this suggests that current tax reform proposals may not produce the economic benefits that their advocates expect.

Cutting tax rates does not increase tax revenues. While perhaps self-evident to many of us raised on the macroeconomics of the 1960's and 1970's, this has become an issue in the 1980's. There is no evidence that the multipliers associated with reductions in taxes are sufficiently large to generate higher revenues (the tax multiplier would have to be at least five).

High and rising structural budget deficits do raise interest rates above levels that would otherwise occur, especially during economic recovery periods. By virtually every measure, real interest rates, especially for long-term instruments, have been extremely high since 1981. Our work at Chase Econometrics indicates that this has in part resulted from deregulation and new innovation in financial markets, which have increased the average cost of funds, increased alternatives to savers, and increased competition for funds. However, the major factor has been the impact of rising actual and anticipated structural budget deficits. We believe that private credit demands are highly income elastic, and thus rise sharply during recovery periods, re-

flecting strong demand for housing, consumer durables, capital goods, and inventories, all of which are heavily financed by borrowing. Rising deficits collide with increases in private credit demands at such times and cause upward pressure on interest rates (if the Fed does not fully accommodate them) or, eventually, will cause upward pressure on interest rates due to higher inflationary expectations (if they are financed by continued rapid growth in the money supply).

There are many who claim that the link between interest rates and deficits is very weak, and that, in fact, interest rates frequently decline when deficits rise. This pattern did indeed occur during 1982. However, this correlation reflects the historical pattern of rising deficits during recessions, when private credit demands have been extremely weak—it is easy to accommodate high deficits during such periods. The problem in more recent years was that deficits rose at the same time that private credit demands increased. This pattern is unprecedented, so that the weak historical correlation between interest rates and budget deficits is essentially irrelevant. Furthermore, the decline in interest rates which occurred in 1982 took place primarily because the Federal Reserve eased dramatically during the second half of that year, plus the fact that the Congress passed legislation which reduced future deficits. The impact of expected future deficits on interest rates in recent years is probably best observed in the relatively steep yield curve which existed during much of that period. Furthermore, enactment of the balanced budget amendment during the latter stages of 1985 coincided with a very significant decline in real long-term interest rates, and a flattening in the yield curve.

Fiscal policy matters and lags are much shorter than those associated with changes in interest rates and exchange rates. The strong recovery in 1983 and early 1984 led to unrealistic expectations about future economic growth because such expectations ignored the impact of excessive deficits on interest rates and the dollar, and the fact that the lags associated with those effects are relatively long. Eventually the adverse effects of those factors began to slow the economy dramatically, coinciding with a fading out of the direct impact of the fiscal stimulus. Gradual reductions in future budget deficits are thus likely to slow the economy in the short run, even though they may improve long-term economic performance, unless financial markets anticipate the effect on interest and exchange rates far in advance, and/or the Fed eases in advance of actual budget cuts or tax increases. With respect to the long term, recent performance suggests that a gradual reduction in future deficits can potentially have a positive impact on long-term economic performance by producing sharp declines in interest rates, as well as in the value of the U.S. dollar on foreign exchange markets, especially if accompanied by a more accommodative monetary policy. Thus, some modest tax increases, as part of an overall deficit reduction package, can be stimulative in the long run, especially if they are phased in slowly.

II. Real Interest Rates

We have already alluded to the high real interest rates which have prevailed in the United States during the early 1980's. This phenomenon has not been unique and by virtually any measure real rates at both the short and long end of the market rose steadily from mid-1980 to early 1982 in virtually all world financial markets. In the United States, Italy, Canada, and the United Kingdom, rates rose to levels well beyond any experience of the postwar period. While declines have occurred in some markets, they remain near record levels in most major markets.

We can speculate about the causes of this phenomenon, and as indicated above, we concur with the general view that the mix of fiscal and monetary policy in the United States has been a contributing factor; however, we believe that there are few who can claim that they anticipated the general level of real rates and none who can claim to have foreseen that these levels of interest costs could have been successfully combined with at least a temporary recovery of growth and investment in the industrialized world.

The surprise is twofold: that real interest rates have remained so high; and that investment and other interest sensitive expendi-

tures have risen, substantially in some cases, despite these high rates. The message for us has at least two dimensions.

Interest elasticity of expenditure and saving to rising interest rates is probably lower than historical estimates suggest. The sample period for estimating these elasticities is dominated by financial regulation which, if not a historical artifact already, is likely to be in the near future. These regulations undoubtedly resulted in corner solutions to market equilibrium and credit rationing that is difficult to identify statistically. We speculate that the coincidence of credit rationing with rising interest rates has resulted in overestimates of the elasticity of demand based on historical data, particularly during periods of rising rates.

There is a hysteresis effect of interest rates on capital expenditures. The impact of any particular level of interest rates on expenditure will depend on the time path of rates. Moreover, it seems likely that there is an essential asymmetry in the response to rates based on stock effects that is not captured in most analyses.

High real capital costs for an extended period of time allow a stock of unexecuted investment projects with returns below the prevailing rate to pile up and essentially pivot the marginal efficiency of capital curve outward at the prevailing rate. At each stage of the decline of rates, that bulge will appear as a one-time increment to investment resulting in a kink in the curve that is likely to be modest at low real rates of interest but may be great when high rates have been maintained for some time. Thus, the direction of change of real rates is likely to be as important as the level in determining the level of demand. If this is correct, a continued gradual decline in rates could extend the present recovery for some time, albeit at a modest pace. If this phenomenon exists, the interest elasticity of demand is likely to be greater in a downward direction than in an upward direction.

III. The U.S. Dollar

From the point of view of the late 1970's, the rise in the value of the dollar which occurred during the first half of the decade of the 1980's must be regarded as one of the outstanding surprises of the period.

Depending on the choice of weights, time period, inflation adjustment, etc., the dollar rose by 60 to 80 percent or more during the 6 years from 1979 to early 1985, and rose fairly steadily. From the perspective of the free-fall of the U.S. currency which occurred during the second half of 1978 and 1979, and the virtually uninterrupted 10-year decline following the floating of the currency in 1971, any outlook incorporating the currency behavior we have experienced would have been audacious indeed. If the same forecast had anticipated that this currency value would be maintained, and indeed increased, in the face of massive and rising U.S. current account deficits of the magnitude we have seen, who of us would have been able to resist challenging the basic consistency of that outlook. Moreover, it appears that the received literature of exchange rate determination is incapable of explaining the shift in currency values, whether it is based on purchasing power parity, monetary demand/supply, or international capital flows and current account balances.

The obvious message from this experience is a fairly lengthy research agenda, but again, we will offer some speculation regarding the results of that research.

Much of the dollar move in the early 1980's was in response to fundamental portfolio adjustments which were freed from constraint. The move of financial deregulation which began on a global basis in the late 1970's has allowed currency shifts in portfolios which were clearly constrained heretofore. The shifts were not just at the margin. In the case of Japan and the United Kingdom, and now potentially France and Italy, fundamental institutional changes unleashed a wave of capital movement to the United States. This would probably have occurred at the margin in any case, but yield differentials and risk perceptions clearly sped the adjustment process.

The Mundell/Fleming thesis on export crowding out will be confirmed for the U.S. economy. Casual observation suggests that the fiscal stimulus to the U.S. economy had a much more serious impact on crowding out exports via exchange rates than in crowding

out investment via interest rates in the face of an only modestly accommodative monetary policy. The high dollar was supported by public and private credit demands which supported higher rates in the United States than abroad in the face of capital market elasticities which have probably risen as a result of financial market deregulation and the rapid rise in the size of international capital flows relative to trade flows and domestic financial markets.

Floating exchange rates are neither the panacea extolled by some nor the impediment to trade and growth challenged by others. The global trading community has already moved well along the road to financial instruments which allow them to shift or accept floating rate risk consistent with corporate risk strategies. In the absence of floating exchange rates, the threat to the world trading environment from U.S. protectionism would not have emerged but a different set of problems associated with U.S. policy mix would confront us. Free exchange markets do shift the risk burden, and may even result in greater overall volatility, than markets under constraint, but the arguments for efficiency and allocation are valid. Moreover, by allowing external producers to buffer the demand increases in U.S. markets, volatility in exchange markets may successfully buffer the activity cycle in the United States allowing a more extended and shallower cycle. (Could the 1981–84 cycle have been even sharper without a floating currency?)

IV. World Oil Prices

The surprise is that world oil prices are not at $50, $60, or $100 per barrel, depending on your favorite forecast at the end of the 1970's. Even if we had forecast the global recession and modest average rate of growth of the global economy, the fall in both the absolute and relative price of petroleum was far from the thoughts of most of us following the 1980 price increase. There are obvious messages from the experience of the 1980s, many of which we have already internalized.

The long-run price elasticity of demand for petroleum is substantially higher than the short-run elasticity. This is likely to be true of any factor of production when demands are tied to large existing capital stocks.

Long-run price elasticities of supply are higher than we had anticipated. We can get more of virtually everything given a high enough price and sufficient time.

Market cartels are likely to be inherently unstable even if relatively long lived. Perhaps no comment is necessary.

In the long run, relative price increases only raise the general level of inflation if absolute price increases are validated by monetary policy. The fundamental difference between the first and second oil shocks was in monetary policy. In the first instance, U.S. monetary policy accommodated a general price increase and inflated away much of the shift in terms of trade. Following the 1980 oil shock, policy failed to validate the inflationary pulse and real incomes were forced to absorb the shift in U.S. terms of trade.

V. Summary

While there have been basic surprises in policy decisions during the first half of the 1980's, most of the unexpected phenomena flowing from those decisions, or associated with them, appear to have been associated with key demand and supply elasticities. We have been through a period during which basic price relatives, including shifts in nominal and real interest rates, the value of the dollar and relative factor prices, moved dramatically relative to historic means and our postwar sample period. Some of the short-run responses have been less than many anticipated while long-run elasticities in several cases have been substantially larger. If we have learned nothing else from this period, we should come away with a healthy skepticism concerning our ability to specify and estimate the dynamics of key product and factor markets.

Supply-Side Modeling from Bits and Pieces

By GEORGE M. VON FURSTENBERG AND R. JEFFERY GREEN*

As the history of Keynesianism illustrates, the distinguishing features and exact claims associated with a particular economic policy doctrine are rarely clear in the early stages of its development. Supply-side economics is no different. Like Keynesianism it has received political applications before its content was settled. Thus, expository work remains to be done before supply-side economics is likely to attract the detailed attention and criticism that will ultimately define its place in the policy discussions of our discipline.

To help place the new approach to economic policy, this paper selects some properties and conditions that could be consonant with major aspects of supply-side doctrine. In preparing this selection, we are quite indifferent to whether those from whose thoughts or models we pick are professed supply siders. Furthermore only two of several possible points of distinction are developed.

The first theme we wish to elucidate is that supply siders insist, more than most economists, that policy should be conducted by rules relating not to procedure (stable money growth), but objectives (stable prices). Procedural rules (tax indexing) or constitutional restraints (balanced budget) may be resorted to as a second-best. Such positions may be advocated if the feedback control mechanisms involved in gearing economic policy instruments directly to final targets are prevented from functioning reliably in the existing political system. In general, however, supply-side economics presents itself as firm and principled on targets, but flexible on methods. The second point is that, to appreciate the elasticity optimism which brought supply-side economics to public attention in the first place, the short run should be extended to one generation and the long run to several generations. The overall impression we hope to convey is that even though supply-side economics has lacked a dominant intellect as challenging as Keynes was to Keynesians and non-Keynesians alike, it may have something beyond mere variations in emphasis to contribute to the concerns of economic policy.

*Department of Economics, Indiana University, Bloomington, IN, 47405.

I. Policy Rules and Intentions

In the formulation of economic policies, supply siders base themselves on preconceived, and potentially constitutionally mandated, final targets. Their pursuit is to be flexible to guarantee success no matter what the slippage in the relation between particular policy instruments and chosen goals. We shall use the supply-side approach to monetary policy to show that a fixed resolve extends only to ends, and that market signals are monitored to adjust the means as necessary to attain those ends. The assumed ability to direct sufficient means at an objective to achieve given ends gives supply siders the confidence to reason back from the final outcome long before it has emerged as desired. By outlining their approach to fiscal policy later in this section, we will show how the interpretation of the current stance differs depending on whether it is projected from past experience or deduced from presumptive achievements of the future.

In characteristic fashion, supply siders define the appropriate monetary policy by its results, and not by the geometry of money supply growth. For them the appropriate money supply rule has to be consistent with the avoidance of inflation while fully accommodating faster growth in potential output stimulated by other means. One way of meeting these requirements is to use an index of sensitive auction-market prices, such as the price of gold, to gauge inflation prospects. As long as that indicator keeps changing, money supply growth would be adjusted in the opposite direction. In this way the authorities would lean against *expected* price changes,

not *past* price changes. Due to staggered contracts and order backlogs, price changes frequently percolate with delay into the official statistics, so that reacting to reported changes in the general price level would not be prompt enough.

Supply siders agree with monetarists that money growth is technically controllable within limits much narrower than the range of variation observed. Hence noise in money growth should be reduced to lower risks in investment planning and production. However, supply siders are far less likely than monetarists to believe 1) that the opportunity cost of money used in transactions is a stable fraction of any interest rate or set of interest rates in a financially deregulated environment, and 2) that convenience in transactions represents almost the entire compensation for forgoing any interest earnings so that anonymity services desired for concealment do not explain much of the demand for currency. More generally, supply siders do not believe that past relationships can be conclusive for determining how much growth in the money supply is consistent with the avoidance of future inflation and deflation, particularly when new supply-side impulses have been given. Rather, for them money growth should be conditional upon current and publicly available market data, and hence subject to modification by events.

To the extent the estimated signal to noise ratio in gold price movements, for general inflation, determined the optimal rate of adjustment in money growth, some reliance would continue to be placed in the stability of past patterns. Overall, however, the goal-seeking policies favored by supply siders ideally should make the U.S. price level (of gold, at least) approximate a random walk without drift (Robert Mundell, 1983). Whether the approach advocated by supply siders would have been consistent with the avoidance of general inflation may soon become apparent. During recent years of declining gold prices, they viewed money growth at least as high as actually occurred as noninflationary, while monetarists continued to warn of its delayed inflationary potential. If inflation rates should rise, it will be difficult to deny that with faster money growth they would have increased more. On the other hand, if inflation does not climb appreciably rather soon, supply siders will remain free to claim that, with even faster money growth, aggregate supply and productive capacity, and not the price level, would have risen more than they did over the past three years.

Although any policy that affects exchange rates may also affect the U.S. dollar price of gold, supply siders view monetary and fiscal policy as having essentially separate allocative tasks. Hence these policies should not be traded off or compromise each other. Discussions of the appropriate policy mix are pointless; rather each policy should be set appropriate to its task. In the fiscal area, supply siders emphasize that disincentive effects rise at an increasing rate with marginal tax rates and with the dispersion of such rates (Victor Canto et al., 1983). Taxing all income earned during an assessment period at the same flat rate could be one part of the answer (Robert Hall and Alvin Rabushka, 1983).

The same concern with minimizing excess burdens, though unadorned with supply-side embellishments, has been used by Robert Barro (1979) to establish intertemporal smoothing of the tax rate as the principle that is to govern changes in the national debt. According to this microeconomic principle of stabilization, any fiscal disturbance should lead to exactly that change in the tax rate that would obviate the need for future tax rate changes, barring further unexpected events. If the ratio of government debt to the national product is ultimately to be maintained at its present level, then any change in the expected present value of government expenditures must be matched by an equal change in the present value of tax receipts produced by a once-and-for-all change in the tax rate. Hence all tax rate changes are to be of equal permanence, but smaller the shorter the expected duration of any fiscal dislocation. If the intended and expected dislocation is permanent, for instance, because the inherited ratio of government spending to *GNP* is viewed as allocatively excessive, taxes should be cut immediately to the long-run sustainable level. That level is lower the greater the expenditure compression ultimately intended, but

higher the longer it is expected to take to get spending down to its final target ratio.

If miscalculations are made in this last regard, and progress toward expenditure reduction is slower than expected, the application of the same principle of smoothing would call for tax rates being raised, as soon as these forecast errors are apparent, to avoid larger tax increases later on. However, as long as a program of reducing the share of government spending in *GNP* is credible, the mere existence of that program would justify an initial string of fiscal deficits which would not have been justified had the old spending trends and practices been left unchallenged. From a supply-side viewpoint confident of eventual success, such deficits would not be structural but transitional.

II. Elasticity Takes Time

Reducing tax disincentives to labor supply, saving, and enterprise is central to the supply-side aspiration of raising, for a time, the rate of growth of potential output (Canto et al.). Changes in tax rates have present and future income and substitution effects. Supply siders claim particular knowledge or appreciation of the size and significance of the latter, without, however, heeding the time that must be allowed for such effects to add up. Both the immediacy of these claims and widespread disillusionment with near-term elasticity optimism are expressed in the following quote from Lyle Gramley:

> When Federal tax rates were reduced dramatically a few years ago, it was hoped that the incentive effects would increase both the rate of personal saving and the willingness to work. Consequently aggregate supply would increase along with the rise in aggregate demand. That was a theoretical possibility, but it did not happen. The personal saving rate has declined, not risen and the increase in the civilian labor force last year was the smallest in two decades. [1984, p. 2]

We shall not deal with tax effects on labor supply and why they may be slow to materialize, as considered, for instance, by Barry Bosworth (1984). Rather we wish to make a point with regard to the response of saving. Here supply-side economics had reason only to assume that, in the long run, we shall all have saved more privately the lighter the taxation of saving. Obviously a rise in private saving need not translate into a rise in national saving unless government deficits prove as transitional as supply siders expect. Furthermore, what happens to the share of *GNP* devoted to domestic capital formation is determined at the margin not so much by national saving as by incentives for foreign and domestic residents to invest in the United States (von Furstenberg, 1983). Hence the bridge between private saving and domestic capital formation has several spans and diversions. Nevertheless, we will focus only on the incentives for private saving. To permit such a restricted view, we assume that the change in private saving correctly measures the change in private net worth resulting from a particular action because no capital gains or losses were induced and no change in future tax liabilities accrued over the period.

For the purpose of analyzing the response speed of saving to tax reduction, Martin Feldstein's (1978) model can be extended from the two periods, working life and retirement, in which consumption C_1 and C_2 occurs, to a third, terminal period. In it bequests are left by the retired and estates settled. The net-of-tax estate, B, reappears as inheritances, I, together with fixed labor income, Y, in the budget constraint of the working generation. If the after-tax nominal interest rate differs from the after-tax real rate, r, by the expected rate of inflation, the present value of the price of retirement consumption relative to the price of current consumption is e^{-r}. Similarly, the relative price of bequests is $(1-u)^{-1}e^{-2r}$, where u is the estate tax rate.

We now attribute to supply siders the belief that the uncompensated price elasticities of demand for retirement consumption and bequests are at least equal to unity. Then, for the lower limit of these values, the utility function, U, may be written with equal (logarithmic) weights, which need not be normalized for the applications that follow, as

$$U = C_1 C_2 B. \tag{1}$$

Maximizing (1) subject to the lifetime budget constraint,

$$(2)\quad Y+I=C_1+e^{-r}C_2+e^{-2r}B/(1-u),$$

yields the condition that the present value of spending, including any taxes, planned for each of the three periods must be the same, that is,

$$(3)\quad C_1=e^{-r}C_2=e^{-2r}B/(1-u)=(Y+I)/3.$$

The total stock of private savings outstanding at any time is then equal to the wealth, W_1, of the working generation and W_2 of the retired generation, such that

$$(4)\quad W=W_1+W_2=(Y+I-C_1)$$
$$+(e^rW_1-C_2)=[1+(e^r-1)/3](Y+I).$$

Wealth, W_3, left after retirement, estate taxes, and bequests must necessarily be zero, so that

$$(5)\quad W_3=e^rW_2-B/(1-u)$$
$$=e^{2r}(Y+I)/3-B/(1-u)=0.$$

Since r is the *ex post* intergenerational real after-tax rate of return, given Y and I, r and hence W rise more, the greater the part of a generation for which taxes on the return to saving have been maintained at a lower level than before. The short run ends after one generation in the three-period model because I can be treated as pregiven only with respect to unexpected tax rate changes which have persisted for no more than one generation. During that interval, the effect of a tax-induced change in r, dr, on total wealth W rises from 0 to $dre^r(Y+I)/3$. The endowment $(Y+I)$ can be expressed in terms of W, using (4). This shows that, even after one generation, the semielasticity of wealth with respect to the compounding factor e^r is still low. It is one-half or less if r is unrestricted, and one-third or less if r is zero or positive in the equation

$$(6)\quad (dW/W)/(dre^r)=(2+e^r)^{-1}.$$

In the long run, however, I and B will converge in an economy with stable tastes which is assumed to return to stationarity in per capita terms. As generations pass after the permanent tax reduction, the bequests which members of the working generation intend to leave will eventually be no higher than the inheritances they have already received from the generation of their grandparents. Replacing I in equation (5) with B and expressing B as a function of r and the fixed value of Y before substituting in (4) gives the long-run (starred) solution for the level of private wealth in the economy,

$$(4')\quad W^*=(2+e^r)Y/[3-(1-u)e^{2r}].$$

Differentiating (4′) and using it to substitute for Y yields the long-run version of the semielasticity below. It can be seen to be the sum of the short-run semielasticity and of another, normally larger, term in the expression

$$(6')\quad (dW^*/W^*)/(dre^r)=(2+e^r)^{-1}$$
$$+2(1-u)e^r[3-(1-u)e^{2r}]^{-1}.$$

For instance, if estate taxes are ignored ($u=0$) and the real after-tax rate of return is assumed to be zero—a value that may have been approximately correct for one or two postwar generations of savers—the second term adds 1 to the short-run semielasticity of one-third. While changes in the taxation of the return on saving have no effect on saving out of labor income, they have some effect on total wealth within a generation and a much larger effect over several generations in the limiting case here considered.

III. Politics and Performance

When supply-side economics spills into politics, it often appears to assert that long-run effects will materialize quickly. The above example may have shown that immediate gratification is not what supply-side economics should look to. Rather, increased regard for long-term relative to short-term benefits would be consistent with the focus on final outcomes and on greater time consistency of

optimal plans described at the beginning of this paper. In fact, however, no reduction in the positive time preference rate used to formulate government policies has occurred, and real interest rates have been higher than in previous decades.

Instead of patiently nurturing the supply side of the economy to earn faster growth in living standards, political practice has encouraged the premature grasping for consumption benefits, no matter how borrowed. Two miscalculations may have contributed to this outcome. In casting about for ways to arrest the growth of government, the Reagan Administration opted for the strategy of cutting taxes first. The hope was that government expenditure reductions and faster economic growth would follow soon after and eliminate the increase in the fiscal deficit resulting from the statutory tax cuts.

Regarding the first point of hope, the ploy of cutting taxes first and then starting to play on fears of large fiscal deficits has not worked to get spending down. Politicians and budget managers continue to promise ever more solemnly that it will be made to work in future years. So far, however, only grants-in-aid and transfer payments have grown less than *GNP*; net interest paid and defense expenditures have grown considerably more. The net result has been a rise in the ratio of U.S. federal government expenditures to *GNP* over the first half of this decade, with state and local governments providing what fiscal austerity there was.

Partly because supply-side economics had emphasized the disincentive effects of high marginal tax rates, there was a second reason for hoping for a quick return to macroeconomic balance when these rates were reduced. If aggregate supply should rise sufficiently on account of the cut in tax rates, domestic absorption, both private and public, could increase without a string of current account deficits whose portfolio consequences would eventually spell trouble. The hope of avoiding instability in external price, trade, and investment relations was disappointed when the United States moved from a position of approximate balance in its current account to one of annual deficits equal to 2 to 3 percent of *GNP*. At an annual rate of $100 billion, these deficits reduce the ratio of U.S. assets abroad to foreign assets in the United States by about 10 percent a year. Had this turn toward negative net foreign investment been accompanied by higher domestic investment rates, the structural external deficits might have been sustainable. As it is, they are not likely to be.

Some years into the new economic policy environment, matters have been left hanging precariously. A sharp decline in the foreign value of the dollar would bring home the delayed political and economic costs of the world's key currency country rapidly increasing its dependence on capital imports. No matter how opportune such capital imports appear to be for the rest of the world when they first get started, and no matter how much they reduce crowding out of private investment by government deficits in the United States initially, they do not permit adjustment to be deferred indefinitely. The desired composition of international portfolios is unlikely to continue to shift so as to accommodate a large *ex ante* excess of domestic investment over national saving in the United States without major changes in interest and exchange rates. When such changes occur, they could reveal the inflationary implications of recent years of historically high money growth. Alternatively, an increase in inflation could trigger a real depreciation of the dollar. Perhaps nothing but the appearance of these medium-term consequences can prod politicians to agree on major reductions in structural fiscal deficits and in money growth. If so, supply-side politics, by piling up problems of austerity for tomorrow in return for short-lived boosts to domestic absorption today, would have revealed itself the enemy of supply-side economics. The latter's major contribution was to have been to shift emphasis in policymaking from short-term expedients to minding the economic base.

Even if supply-side economics has survived the translation from economics to politics no better than other economic policy doctrines, it cannot be divorced from politics. Supply-side economics, like Keynesianism, is directed at political application; it must therefore continue to be taken up by politicians to retain the relevance it has aimed for. If the push of supply-side politics continues, it may

give rise to much further reorientation of economic policies throughout the world. For this reason, there is continuing interest in its bits and pieces, and what economic whole they may make beyond present failings in design and execution.

One may suspect that these failings are due not to the political equivalent of random error but to deeply held erroneous suppositions about stylized facts and political processes. As detailed in our articles with Jin-Ho Jeong (1985, 1986), the assumption of "tax and spend," which, judged by historical experience, overestimates the leverage of taxes on government spending, may have contributed to the wide gap between promises and performance in the fiscal arena. Undue elasticity optimism may also have contributed to untoward developments such as continued growth in the size of government in relation to *GNP* and persistence of large deficits into advanced stages of expansion. It remains to be seen whether the flexibility in the choice and use of means brought by supply-side economics to the pursuit of economic policy objectives also extends to the adjustment of preconceptions as experience may require.

REFERENCES

Barro, Robert J., "On the Determination of the Public Debt," *Journal of Political Economy*, October 1979, *81*, 940–71.

Bosworth, Barry P., *Tax Incentives and Economic Growth*, Washington: The Brookings Institution, 1984.

Canto, Victor A., Joines, Douglas H. and Laffer, Arthur B., *Foundations of Supply-Side Economics: Theory and Evidence*, New York: Academic Press, 1983.

Feldstein, Martin, "The Rate of Return, Taxation, and Personal Savings," *Economic Journal*, September 1978, *88*, 482–87.

Gramley, Lyle E., "Can the Recovery Be Sustained?," *The MGIC Newsletter*, May/June 1984, 1–3.

Hall, Robert E. and Rabushka, Alvin, *Low Tax, Simple Tax, Flat Tax*, New York: McGraw-Hill, 1983.

Mundell, Robert A., "International Monetary Reform: The Optimal Mix in Big Countries," in J. Tobin, ed., *Macroeconomics, Prices, and Quantities*, Washington: The Brookings Institution, 1983, 285–95.

von Furstenberg, George M., "Domestic Determinants of the Current Account Balance of the United States," *Quarterly Journal of Economics*, August 1983, *98*, 401–25.

______, **Green, R. Jeffery and Jeong, Jin-Ho**, "Have Taxes Led Government Expenditures: The United States as a Test Case?," *Journal of Public Policy*, No. 3, 1985, *5*, 321–48.

______, ______, **and** ______, "Tax and Spend, or Spend and Tax?," *Review of Economics and Statistics*, forthcoming, 1986, *68*.

Sex Segregation Within Occupations

By WILLIAM T. BIELBY AND JAMES N. BARON*

Do the paychecks and career prospects of men and women differ because of rational, optimizing decisions made by men and women, given their respective skills, experiences, home responsibilities, and intentions regarding future work and family activities? The human capital model—coherent, logical and internally consistent—has a certain elegance to it. Nevertheless, empirically, it cannot explain the sex composition of jobs in lines of work pursued by both men and women. The results reported below are more consistent with the notion that employers statistically discriminate; that is, they reserve some jobs for men and others for women based on their perceptions of the average traits of the two groups. However, our results also call into question the assumption that statistical discrimination is merely an optimal decision-making strategy under conditions of uncertainty.

We summarize below our recent research on determinants of the sex composition of jobs in mixed occupations. That research is based on data on staffing patterns, job requirements, and personnel policies collected by the U.S. Employment Service in over 400 California establishments between 1959 and 1979. Details on sample representativeness, procedures of data collection, and the like can be found in our 1984 and 1986 papers.

†*Discussants*: Heidi I. Hartmann, National Academy of Sciences; Glen C. Cain, University of Wisconsin.

*Professor, Department of Sociology, University of California, Santa Barbara, CA 93106, and Assistant Professor, Graduate School of Business, Stanford University, Stanford, CA 94305. This research was supported in part by a grant from the National Science Foundation (SES 79-24905) and research funds from the Academic Senate, University California-Santa Barbara, and the Graduate School of Business, Stanford University.

I. Job Segregation and Sex Differences at Work

Job segregation is the proximate cause of sex differences in earnings and career prospects. For a variety of reasons, men and women end up working in different occupations. However, considerable sex segregation occurs within detailed occupational categories as well. For example, men and women in the same occupation often work for different employers and, more often than not, the employers of men provide better work opportunities. Moreover, even when men and women in the same occupation work for the same employer, they may be assigned to distinct job titles with different duties, responsibilities, and opportunities for training and advancement. Indeed, we found that only 10 percent of the nearly 61,000 workers in our sample of establishments were in job titles that had both men and women assigned to them (see our earlier papers). Furthermore, sex segregation was nearly as extreme when we excluded from our analysis jobs in detailed occupations that were more than 80 percent male or more than 80 percent female. Consequently, sex segregation—by establishment or by job title within establishments—is almost always the mechanism generating sex differences in pay, promotion opportunities, and other career outcomes. In our sample of establishments, unequal pay or promotion prospects for the same job assignment were rarely an issue, since men and women seldom shared the same job title in an establishment. Moreover, this was true even when the work of men and women is classified into the same detailed occupation.

II. Human Capital Theory

Human capital theory does not adequately account for sex segregation in job assignments. We have argued that job segregation

by sex is the *proximate* cause of gender differences in socioeconomic opportunities. It could still be the case that human capital theory explains why men and women end up in different firms or in distinct job titles in the same firm. For example, men and women may choose to specialize in different jobs. Even within the same detailed occupation, it might be easier to combine work and family responsibilities in some jobs than in others. Or, even if men and women do not deliberately choose different firms or job assignments, their experiences and training might not qualify them for the same jobs, even when they are in the same general line of work.

To test the human capital explanation, we examined how men and women in the same general line of work get distributed across specific job titles within establishments. First, we identified detailed occupations that were no more than 80 percent male or female in our sample. Then we examined the determinants of sex composition for establishment job titles that were classified into one of those mixed occupations. The human capital model emphasizes the importance of training, skills, and turnover costs in determining the job assignments of men and women. To test the model, we examined the extent to which the sex composition of jobs was determined by measures of those factors as ascertained by Employment Service job analysts. The specific measures are listed in Table 1 (complete details appear in our 1986 paper). The model represents the determinants of percent female in the ith job in occupation j within establishment k as

$$(1)\quad P_{ijk} = a + b_1 x_{ijk} + b_2 z_k + u_j + e_{ijk},$$

where x_{ijk} is a vector of job attributes, z_k is a vector of establishment characteristics, u_j is a term representing all unmeasured occupational traits that affect the sex composition of jobs in a line of work, and e_{ijk} is an orthogonal job-specific stochastic disturbance.

Human capital variables may explain why some occupations are predominately male while others are overwhelmingly staffed by females. However, our analyses revealed that skills, training, and turnover costs have a remarkably weak impact on how men and women in *mixed* occupations get distributed across organizations and across job titles within establishments. Our statistical model explained just 14 percent of the variance in the sex composition across establishments for men and women in the same occupation and only 16 percent of the variance in the sex composition of specific titles in the same occupation within establishments. Under the model, nearly two-thirds (63 percent) of the jobs in our sample are predicted to have between 20 and 80 percent females. However, only 6 percent actually did; 54 percent of the jobs from mixed occupations contained no women, and 39 percent included no men. In short, a remarkably rich collection of measures of training and human capital requirements of jobs failed to explain the degree to which women have access to the same jobs as men in mixed lines of work.

TABLE 1—LOGISTIC REGRESSION COEFFICIENTS FOR LIKELIHOOD THAT WOMEN ARE EXCLUDED FROM A JOB IN A MIXED OCCUPATION ($N = 2997$)

Independent Variable (and Metric)	Coefficient[a]
z_1 Organizational Scale (log employment)	.20[d]
z_2 Union or Bidding Arrangements (0-1)	.59[d]
x_1 Specialization[b] (log workers)	.60[d]
x_2 Training Time (1-7)	.29[d]
x_3 Numerical Aptitude (1-4)	.33[d]
x_4 Verbal Aptitude (1-4)	−.70[d]
x_5 Finger Dexterity Aptitude (1-4)	−1.13[d]
x_6 Clerical Perception (1-4)	−.58[d]
x_7 Spatial Skill (1-4)	.52[d]
x_8 Eye/Hand/Foot Coordination (1-4)	.52[d]
x_9 Physical Strength[c] (0-1)	1.45[d]
x_{10} Varied Duties (0-1)	−.27
x_{11} Repetitiveness (0-1)	−.39[d]

Note: Likelihood Ratio *Chi*-square = 1001.4 with 22 degrees of freedom.

[a]Maximum likelihood estimate.

[b]Sign reversed so that high scores correspond to specialized jobs.

[c]Lifting 25 pounds or more.

[d]$p < .001$.

III. Statistical Discrimination

Job segregation within mixed occupations appears to be generated by statistical discrimination by employers. The same set of variables allowed us to determine quite con-

clusively whether or not a woman would be *excluded* altogether from a given job title in a mixed occupation. Respecifying equation (1) as a logistic regression of the odds of exclusion allowed us to correctly predict whether or not women were excluded from jobs in mixed occupations for 82 percent of the cases. Moreover, logit coefficients, reported in Table 1, show that two measures—Employment Service analysts' ratings of jobs' *physical demands* and requirements for *finger dexterity*—had by far the greatest impact on whether women are excluded from a job. Other measures of training and human capital requirements were of secondary importance. (See our 1986 paper for a detailed discussion.) Our results suggest that decisions about allocating men and women to jobs in mixed occupations are strongly influenced by characteristics of job candidates presumed to vary by sex. If, for example, a job is viewed as physically demanding, it is typically deemed inappropriate for all women. Conversely, a job viewed as requiring finger dexterity is typically seen as inappropriate for all men. In short, we have strong evidence that some employers in our sample practiced *statistical discrimination*: they reserved some jobs for men and others for women, based on perceptions of group differences between the sexes.

It is claimed that employers rationally rely on group differences when it is difficult and costly to determine the qualifications and likelihood of turnover for specific job candidates. Our results suggest that statistical discrimination is neither particularly rational nor efficient. Variables most directly related to the turnover costs of jobs had only small effects in our statistical model. The variables that did distinguish between male and female jobs (in occupations pursued by both men and women), finger dexterity and the ability to lift at least 25 pounds, are relatively easy to assess for individual men and women. In short, it is difficult to imagine how sex operates as an efficient "screen" utilized by rational, unbiased, profit-maximizing employers. Instead, it seems that stereotypical notions of "men's work" and "women's work" play a more decisive role than did the technical qualifications of individual applicants and the skill requirements of jobs.

IV. Employer Policy

Sex segregation is imbedded in organizational policies and sustained by organizational inertia. The impact of physical demands showed up in our qualitative analyses of organizational policies as well. Our evidence suggests that the statistical effects described above came about because of employers' deliberate policies restricting women's access to jobs perceived to be physically demanding. Court decisions in the early 1970's invalidated the legal basis for these policies, and, not surprisingly, we found fewer references to such policies in organizations studied since 1971. However, court decisions barring the use of physical demands as a criterion for excluding women from specific jobs appeared to have little effect on employers' practices in the 1970's. In our statistical model, we allowed the effect of physical demands to differ between establishments studied prior to 1971 and those studied since 1971. The effect of physical demands on the likelihood of excluding women from a job was actually slightly larger for the organizations studies more recently (see our 1986 paper).

References to policies excluding women from jobs perceived to be physically demanding appeared in about 40 percent of the narrative reports prepared by Employment Service analysts. Moreover, these policies were invoked in diverse organizational and industrial contexts (see our 1986 paper for details). Furthermore, such policies were often referenced for jobs that actually required no strenuous physical exertion, according to detailed job analyses.

The basis of sex segregation in the policies and formal arrangements of organizations showed up in other statistical analyses as well. In attempting to explain why some organizations are more segregated than others, we found that small, entrepreneurial firms are often completely segregated by sex. In these establishments, women were either excluded altogether or confined to just one or two job classifications, such as "receptionist." Among the remaining organizations, we found higher levels of segregation in the larger establishments, those with more specialized jobs and a proliferation of job

titles, and in unionized settings and those with formal bidding procedures governing promotions (see our 1984 paper). In short, instead of promoting universalistic standards in personnel matters, bureaucratic rules and procedures seem to have been implemented in a way that sustains job segregation by sex. Moreover, longitudinal analyses of a subset of establishments (those analyzed twice by the Employment Service, typically 4 to 6 years apart) revealed that large, bureaucratic firms are least likely to show change in the level of segregation in the absence of any sustained, deliberate effort to bring women into jobs previously closed to them.

V. The Social Psychology of Perception

Sex segregation is sustained by behavior as well as policy. Economists are certainly correct in insisting that specialization by sex comes about in part because of the choices individual men and women make regarding their work and nonwork roles. How individuals form attitudes and make commitments, perceive others, and choose among alternative activities influences the roles men and women play in the workplace. Unfortunately, economists' models of employers' behavior have been unaffected by recent research on perception, attitude formation, and choice behavior.

For example, social psychologists have gained considerable insight into the process of "person perception." They have discovered that *stereotypes* seem to be an essential feature of human perception, a cognitive shorthand we invoke in order to achieve economy in the processing of information (Richard Ashmore and Frances Del Boca, 1985). This research suggests that, despite what is typically assumed, even employers have limited cognitive capabilities and are not immune from the use of stereotypes.

According to the model of statistical discrimination, employers base decisions on their perceptions of group differences when it is difficult and costly to determine the true qualifications of individual applicants. It is not difficult to see how the phenomenon social psychologists call "expectancy confirmation sequences" (J. M. Darley and R. H. Fazio, 1980) turn this seemingly efficient behavior into a self-fulfilling prophecy. According to this view, employers expect certain behavior from women (for example, high turnover) and therefore assign them to routine tasks and dead-end jobs. Women respond by exhibiting the very behavior employers expect, thereby reinforcing the stereotype. Moreover, individuals are more likely to attend to and retain information that confirms stereotypes and to ignore information that belies their preconceived notions (David Hamilton, 1981). Therefore, it is likely that employers ignore information that fails to fit their expectations regarding men and women workers. Social psychologists consistently elicit and isolate stereotyping and expectancy confirmation processes in their laboratory experiments. These processes certainly operate in the workplace, where long-term interactions among employers and employees allow patterns of behavior similar to those elicited in the laboratory to stabilize and become taken for granted.

VI. Summary and Conclusion

Our research has examined the determinants of the sex composition of jobs in occupations employing both men and women in a diverse sample of California establishments studied by the U.S. Employment Service between 1959 and 1979. We found that when men and women performed similar work roles, their jobs were typically either in different enterprises or in segregated titles within an organization. While skills, aptitudes, training, and turnover costs influenced the sex composition of job assignments, stereotypical traits such as physical demands and finger dexterity had the greatest impact on whether women were excluded from a job in a mixed line of work. Our findings are consistent with the theory of statistical discrimination, which posits that employers reserve some jobs for men and others for women. However, our evidence calls into question the claim that employers' practices reflect efficient and rational responses to sex differences in skills and turnover costs.

We found pervasive job segregation by sex in our sample. That segregation appeared to be rooted in organizational policies and structures which remained stable unless de-

liberate efforts were made to change them. Only recently have we developed alternative theories and accumulated empirical evidence about why work roles are segregated by sex. Our findings lead us to conclude that: much of the disparity between men and women in economic well-being and career prospects is due to segregation at work; and intervention can make the workplace more equitable.

REFERENCES

Ashmore, Richard D. and Del Boca, Frances K., "Gender Stereotypes," in their *The Social Psychology of Female-Male Differences*, New York: Academic Press, 1985.

Bielby, William T. and Baron, James N., "A Women's Place is with Other Women: Sex Segregation Within Organizations," in B. F. Reskin ed., *Sex Segregation in the Workplace: Trends, Explanations, Remedies*, Washington: National Academy Press, 1984.

_____ **and** _____, "Men and Women at Work: Sex Segregation and Statistical Discrimination," *American Journal of Sociology*, January 1986, *91*, 759–99.

Darley, J. M. and Fazio, R. H., "Expectancy Confirmation Sequences," *American Psychologist*, 1980, *35*, 867–81.

Hamilton, David L., *Cognitive Processes in Stereotyping and Intergroup Behavior*, Hillsdale: Erlbaum, 1981.

Internal Labor Markets and Noncompeting Groups

By PETER B. DOERINGER*

The concept of the internal labor market was intended to emphasize firms and unions as the principal institutions that segment the labor market, whereas the idea of the dual labor market was only an allegory for highlighting issues of poverty and discrimination. (See my book with Michael Piore, 1971, 1985). Nevertheless, it was labor market dualism that attracted most of the subsequent attention. This paper draws upon several recent research themes that have revived interest in internal labor markets to illustrate the possibilities for a socioeconomic theory of the labor market—one which emphasizes *collective* instead of *individualistic* behavior, and which leads to the proliferation of noncompeting groups in the labor market.

I. Enterprise-Specific Skills, Implicit Contracts, and Efficiency Wages

The original work on internal labor markets emerged from a series of field interviews with employers and unions. It provided an explanation for the wage rigidity, employment tenure, and establishment wage effects that characterize parts of the labor market, and it also corresponded with more mainstream research on enterprise-specific skills and the economics of on-the-job training (Gary Becker, 1964). Subsequently, various other elements of internal labor markets were incorporated into economic theories of wage and employment determination (Janet Yellen, 1984).

For example, theories of implicit contracts providing for fixed wages and secure employment showed how internal labor markets could be an efficient response to a fluctuating and uncertain economy. By postulating that individual workers were more risk averse than employers, the implicit contract literature defined opportunities for arbitrage between workers and employers in which guarantees of stability of income could be exchanged for lower compensation levels than would apply under a regime of flexible wages and employment.

Efficiency wage strategies, where employers raise wages above competitive market rates in order to promote productivity or deter shirking, provide an explanation of why compensation levels within internal labor markets might not fall to market-clearing levels. By using efficiency wages, employers can compensate for problems of agency, imperfect information, and opportunism in internal labor markets by manipulating the opportunity costs of quits and discharges.

II. Neglected Elements

By focusing on efficiency considerations within internal labor markets, however, attention has been diverted from several less competitive features of such markets—their potential for generating economic privileges and rents, the opportunities they provide for "costless" asset formation, the presence of bargaining power, and the importance of social relations and group cohesion at the workplace. Taking these omissions into account helps to explain why internal labor markets are highly resistant to competitive influences and why monopoly elements can permeate the labor market.

Enterprise-Specific Privileges and Rents: Both enterprise-specific investments and efficiency wage strategies result in increased labor productivity within internal labor markets. Moreover, such practices imply that employers will seek to ensure these productivity gains by voluntarily sharing them with workers so that internal wages will be higher than external market opportunities. These "efficient" rents, however, are only an example of a broader class of privileges and benefits that are available exclusively to

*Professor of Economics, Boston University, Boston, MA 02215. I have benefited from discussions with Steven Stoft, Chris Ruhm, Michael Piore, and Paul Osterman.

"insiders." The presence of such privileges provides the economic glue for sustaining internal labor market arrangements.

Costless Asset Formation: While the privileges needed to sustain internal labor markets can be generated by incurring the costs of training investments or efficiency wage payments, they can also arise from various costless practices. For example, case studies indicate that much on-the-job learning is incidental to working, or even "playing" with machinery. Such learning by osmosis is often neither costly nor part of a deliberate investment decision. Similarly, ideology, loyalty, and friendship can also be costless devices for enhancing productivity (George Akerlof, 1982). It is, therefore, questionable whether many of these productivity-enhancing practices are appropriately characterized by investment processes or other micro-efficiency approaches.

Bargaining Power: Implicit in theories of efficiency wage and of enterprise-specific training is the idea that individual workers acquire the power to penalize employers, power that is not otherwise present in the labor market. Workers with enterprise-specific skills can quit, thereby terminating the returns to training and forcing employers to incur replacement costs. According to efficiency wage theory, employees can engage in shirking behavior that can disrupt production.

These sources of individual bargaining power can be further strengthened by collective action. Shirking by individual workers is easier to monitor and to control through discharges than is *collective* shirking. The penalty from quits by individual workers with specific skills is confined to the replacement costs of those workers, whereas collective quits impose further costs of replacing that part of the firm's on-the-job training capability that is embodied in the incumbent workforce.

This individual and collective bargaining power arises only in work situations that generate productivity-based rents. Therefore, there is a concurrence between bargaining power and something of value to be bargained about. With bargaining power comes the possibility of additional transfers of productivity rents to workers, thereby blocking the competitive dissipation of such rents through expanded employment or lower prices.[1]

Social Relations and Group Cohesion: The creation of economic rents and bargaining power are two elements of internal labor markets; the formation of social norms by stable social groups at the workplace is a third. Where enterprise-specific rents are paid to workers, labor turnover is deterred and bargaining power can be used to prevent terminations. As a result, work groups become highly stable. While the exploration of the social dynamics of stable work groups is beyond the scope of this paper, two characteristics of such groups are worth noting —their effect on productivity and their concern with fairness, legitimacy, and income distribution.

For example, work loads and production speeds are rarely determined unilaterally by management. In unionized plants, they are often the explicit subject of bargaining and work groups appear to be important in setting work norms, even in nonunion situations (Sumner Slichter et al., 1960). In addition, established work groups can affect productivity through control over the assimilation and training of new employees.

Labor productivity is also affected by equity and legitimacy at the workplace. The literature on industrial relations emphasizes unions' interest with fair treatment; the personnel management literature stresses the importance of equitable personnel practices in nonunion firms (Slichter et al.; Fred Foulkes, 1980). Economists have sometimes incorporated this idea of equity to internal wage setting by arguing that there is interdependence among workers' utility functions with respect to the internal structure of wage rates.[2] As a result, the *distribution* of income within stable work groups is no longer incidental to earnings determination, as it is

[1]Notions of individual bargaining power seem to underly Oliver Williamson's (1975) discussion of opportunism by workers. His position on the issues of rents is less clear.

[2]The substance of this notion is developed in my book with Piore, ch. 4. See also, Akerlof.

under labor market competition. Instead, it is an explicit concern of the work group and is subject to the test of legitimacy. (Nicholas Abercrombie et al., 1980).

Feedback Effects: The preceding discussion outlines a model of internal labor markets in which economic rents are generated in ways that endow workers economic power and which lead to the formation of stable work groups. These stable work groups then develop norms regarding the fairness of work effort and the distribution of economic rents. Such a model incorporates the social dimension of internal labor markets and provides an alternative to efficiency-driven explanations of relative wage rigidity, employment tenure, and tenure-wage profiles within internal labor markets.

There are also feedbacks within this model that further contribute to bargaining power and rent generation. For example, stable work groups are more cohesive than transient groups, and are more likely to develop collective sanctions. Collective sanctions will tend to increase the share of economic rents paid to the internal labor force and further contribute to work-group stability and cohesion.

A second example of feedback is in the relationship between legitimacy and productivity. The stable work groups found in internal labor markets apply tests of legitimacy to both compensation and effort. Where there is legitimacy, morale, job satisfaction, and productivity are likely to be enhanced (Richard Freeman and James Medoff, 1984). Where legitimacy is weak, shirking and labor turnover may cut into the economic rents being generated internally.

III. Noncompeting Groups

Efficiency-based theories explain the permanence of internal labor markets in terms of their cost competitiveness while largely neglecting the issues of power, economic rents, and equity and distribution. Incorporating these latter effects into the internal labor market model introduces the possibility of dynamic improvements in labor productivity. Rising labor productivity generates economics rents that can be appropriated by workers (and by employers, where product markets are less than competitive), and creates additional economic incentives for preserving internal labor market arrangements that generate and protect rents. The availability of rents and bargaining power can also explain why specific internal labor market arrangements, once in place, tend to endure in the face of changing competitive conditions.

The conventional internal labor market is, however, only one manifestation of far more general processes that govern the labor market. There is some evidence that a wide range of labor markets actors—employers, unions, professional associations, families, and informal labor market groups—strive to improve their economic position by creating employment situations within which rents are generated and then distributed to the work group.

This generation of rents can take many forms, all of which involve introducing noncompetitive elements into the labor market. Examples of this phenomenon include the attempts of craft unions to establish territoriality over the labor supply; "job ownership" at the workplace which prevents incumbent employees from being underbid by workers outside the firm; and family-based work groups in which entry is limited to family members.

Such monopoly behavior is deliberately intended to differentiate between internal and external labor markets. This may arise as a result of enterprise-specific skills, from efficiency wage strategies, or through classic labor market monopoly. It may also reflect worker solidarity as in craft unions or professional associations; or it may emerge as a "social asset" resulting from cooperation within kinship and friendship-based work groups (N. V. Jagannathan, 1984).

Regardless of the source of internal rents, the consequences are similar. Work effort can be increased; rents become available exclusively to the members of the work group; and nonmarket discretion can control the distribution of income within the group. As a result, static welfare losses from such monopoly behavior can be offset by dynamic welfare gains.

IV. Some Evidence

Analyzing the economic effects of noncompeting groups requires information on both the specific rules that govern such groups and on the *mix* of collective labor market institutions. This information is not available in the usual surveys of worker or establishment characteristics. The empirical basis for these propositions must, therefore, be fragmentary and somewhat circumstantial.

One body of economywide evidence comes from the work of Freeman and Medoff (ch. 11), that emphasizes the positive impact that unions and collective bargaining can have upon labor productivity, partly by promoting legitimacy at the workplace. These findings presumably extend as well to those larger, nonunion firms that emulate unionized workplaces (Foulkes). The literature on the economics of discrimination can also be seen as capturing some of the effects of internal labor markets on income distribution.

A second line of research involves the direct examination of institutions, and their internal rules, through case studies. For example, research on informal sector labor markets in developing countries suggests that, far from being atomistic and unstructured, they are full of noncompeting groups that restrict entry and generate privileges and rents for their members. (See my 1983 report and also Jagannathan.) These include family farms and businesses, craft and artisanry workshops, and even highly informal work groups such as pedicab drivers and cigarette sellers (Gustav Papanek, 1975). Most informal sector workers appear to belong to elaborate social systems for generating what Jagannathan defines as "social assets" (loyalty, patronage, and connections) that yield economic value in factor and product markets. Only those few workers who are totally destitute and cut off from family and social support systems are unable to obtain shelter from competition in rent-generating work groups.

There are also several lines of field research on internal labor markets in industrialized countries that provide examples of the different ways that noncompeting groups and social factors generate and distribute internal rents (Marc Maurice et al., 1984). Research on paternalism in internal labor markets—the practice of building loyalty and fostering individual worker dependence on the employer through discretionary economic benefits—illustrates another alternative to financial incentives as a motivator of labor productivity, one which tends to distribute rents towards employers. Under paternalism, legitimacy is maintained through reciprocal arrangements between the employer and individual workers, whereby economic protection is granted to workers in exchange for allegiance to the employer and a commitment to high levels of effort and accommodation at the workplace (see my 1984 paper).

Several studies have emphasized the importance of kinship in the labor market in a wide range of industries from fishing to apparel. (See my forthcoming article with Philip Moss and David Terkla.) Wages and employment on family-owned and operated fishing vessels, for example, are subject to work- and income-sharing guarantees among family members, in contrast to more capitalistic vessels in which the employment relationship is relatively impersonal, and in which labor input is varied with output. While there are often static inefficiencies caused by overmanning on family vessels, these inefficiencies appear to be more than offset by higher productivity and performance, in comparison to their capitalist counterparts.

VI. Implications for Economic Performance

The title of this paper stresses noncompeting labor market behavior. Workers and employers do not simply take advantage of technical opportunities to capture economic rents in a basically competitive system (Anne Krueger, 1980). Instead, they try to generate rents through monopolization and through various idiosyncratic workplace practices, some of which are costless. In this process, workers acquire bargaining power through the economic sanctions that they can impose upon the rent-generating process, and they also acquire social cohesiveness.

The focus on noncompeting groups is intended to emphasize the importance of collective behavior that segments labor markets. These segments, however, are not totally immune to competition. Labor markets are contestable through entry of new work groups and through competition in the product market. Such competition can be resisted and deferred, but never fully overcome. In contrast to classical monopoly power, however, this analysis of monopoly power can increase productivity and output.

It would be premature to conclude from these sketchy arguments and examples that monopolization of labor markets and collective forms of work organization necessarily represent a superior form of economic organization to the impersonal firm of mainstream micro theory. They do, however, reaffirm that there are social, as well as economic, foundations to internal labor markets. Theories of internal labor markets that focus exclusively on human capital investment processes, price incentives, and efficiency miss a whole class of social considerations that can positively affect productivity by reducing competition in the labor market.

REFERENCES

Abercrombie, Nicholas et. al., *The Dominant Ideology Thesis*, London: Allen & Unwin, 1980.

Akerlof, George A., "Labor Contracts As A Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–70.

Becker, Gary S., *Human Capital: A Theoretical and Empirical Analysis, With Special Reference to Education*, New York: Columbia University Press, 1964.

Doeringer, Peter B., "Segmenting Forces in Labor Markets: Towards a Theory of Work Groups and Employment Systems," Report to the International Labour Organization, mimeo., September 1983.

______, "Internal Labor Markets and Paternalism in Rural Areas" in Paul Osterman, ed., *Internal Labor Markets*, Cambridge: MIT Press, 1984.

______ **and Piore, Michael J.,** *Internal Labor Markets and Manpower Analysis*, Lexington: D.C. Heath, 1971; Armonk: M. E. Sharpe, 1985.

______, **Philip I. Moss, and Terkla, David G.,** "Capitalism and Kinship: Do Institutions Matter in the Labor Market?," *Industrial and Labor Relations Review*, forthcoming.

Foulkes, Fred K., *Personnel Policies in Large Non-Union Companies*, Englewood Cliffs: Prentice-Hall, 1980.

Freeman, Richard B. and Medoff, James L., *What Do Unions Do?*, New York: Basic Books, 1984.

Jagannathan, N. V., "Extra-Legal Property Rights in Production and Labor Markets," unpublished doctoral dissertation, Boston University, 1984.

Krueger, Anne O., "The Political Economy of the Rent-Seeking Society," in James M. Buchanan et al., eds., *Toward A Theory of the Rent-Seeking Society*, College Station: Texas A&M University Press, 1980.

Maurice, Marc et al., "The Search for a Societal Effect in the Production of Company Hierarchy: A Comparison of France and Germany," in Paul Osterman, ed., *Internal Labor Markets*, Cambridge: MIT Press, 1984.

Papanek, Gustav F., "The Poor of Jakarta," *Economic Development and Cultural Change*, October 1975, *24*, 1–17.

Slichter, Sumner H., Livernash, E. Robert and Healy, James J., *The Impact of Collective Bargaining on Management*, Washington: The Brookings Institution, 1960.

Williamson, Oliver E., *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

Yellen Janet L., "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, *74*, 200–05.

Work Power and Earnings of Women and Men

By MARIANNE A. FERBER, CAROLE A. GREEN, AND JOE L. SPAETH*

Numerous studies have established that part of the very substantial male-female earnings gap is explained by differences in the amount of human capital workers have accumulated. (See, for example, Jacob Mincer and Haim Ofek, 1983.) Institutional factors have also been found to play a role in determining wages (David Gordon et al., 1982). Occupation further helped to explain the remaining gap, but several researchers have shown that introducing dimensions of work authority by taking into account the individual's position in the work hierarchy explains more of the variation in earnings than does occupation (Martha Hill, 1980). Last, two recent studies (Ferber and Spaeth, 1984; Spaeth, 1985) also included control over monetary resources. This variable added substantially to the explanatory power of earnings regressions, even after human capital variables, institutional factors, and several other measures of work authority had been entered.

Like the other studies, Ferber and Spaeth also found that reward structures for men and women are quite different, suggesting the possible existence of discrimination. The question whether women may also be at a disadvantage in achieving control over monetary resources was not investigated. When Hill examined the process of achievement of work authority, she found substantial differences between male and female workers. In this paper we examine whether the same is true for attaining financial control.

*Professor, Department of Economics, University of Illinois, Urbana-Champaign, Champaign, IL 61820; Assistant Professor, Department of Economics, University of South Florida; and Professor, Department of Sociology and Research Professor, Survey Research Laboratory, University of Illinois, Urbana-Champaign, respectively. We thank John Jackson for advice on econometric matters, and Barbara F. Reskin for helpful comments.

I. Data and Analysis

The data used for this research were collected in 1982 as part of a practicum in survey research methods by carefully trained students. Telephone interviews were conducted with 245 women and 312 men living in the state of Illinois who were employed at least 20 hours a week on a single job. The sample used in this paper excludes self-employed workers and has a total number of 416. In addition to the usual questions about work history and classifying characteristics, detailed information designed to measure quite specific work characteristics was obtained. A more detailed description of the data collection and construction of variables is provided in Spaeth's earlier article. The measure of monetary control, which is the special focus of this paper, is based on the following three questions: 1) About how much money was involved in the largest monetary decision in which you participated last year? 2) Did you make the final decision, make recommendations or give advice, or provide information? 3) How much money can you spend without getting authorization from higher up?

These data are first used to estimate the effects of monetary control on models containing variables of the kind mentioned above. In order to determine whether there are significant differences between the sexes in the rewards associated with control over monetary resources, these regressions are also run for men and women separately. In addition, we attempt to determine whether women have the same opportunity as men to attain control over monetary resources with the same levels of human capital and in the same institutional setting. Because a large number of members of our sample have no such authority, we use a Tobit procedure to allow for censoring. Finally, we estimate the extent to which women earn less than men when differences in the opportunity to attain

TABLE 1—CHARACTERISTICS OF WORKERS AND THEIR JOBS[a]

	Men's Means	Women's Means
Years of Education	13.62	13.47
	(2.64)	(2.41)
Years of Experience before Current Employer	5.92	5.21
	(8.74)	(7.46)
Years with Current Employer, Previous Job	3.11	1.33
	(6.64)	(3.61)
Years in Current Job	7.98	5.16
	(8.33)	(6.39)
Weeks Worked	47.88	45.34
	(9.37)	(11.60)
Hours Worked	42.80	38.74
	(7.84)	(8.27)
Married (Married = 1)	.70	.49
	(.46)	(.50)
Core Industry	.78	.78
	(.41)	(.42)
Ln Number of Employees	6.64	5.76
	(3.15)	(2.92)
Control over Monetary Resources	37.57	34.48
	(13.67)	(8.85)
Number of People Supervised	4.55	2.91
	(9.80)	(8.55)
Sex of Supervisor (Male = 1)	.94	.56
	(.24)	(.50)

[a]Standard derivations are shown in parentheses.

TABLE 2—DETERMINANTS OF Ln OF EARNINGS ($N = 416$)

Years of Education	0.0603[c]
Years of Experience before Current Employer	0.0198[c]
Years of Experience before Current Employer, Squared	−0.0003
Years with Current Employer, Previous Job	0.0195[a]
Years with Current Employer, Previous Job, Squared	−0.0004
Years in Current Job	0.0171[c]
Weeks Worked	0.0089[c]
Hours Worked	0.0169[c]
Married	0.1034[b]
Core Industry	0.0884
Ln Number of Employees of Firm	0.0289[c]
Control over Monetary Resources	0.0080[c]
Number of People Supervised	0.0035[a]
Sex of Supervisor	0.2470[c]
Sex of Employee	0.2891[c]
Intercept	−0.2185
R^2	.56

[a]Significant at .10 level.
[b]Significant at .05 level.
[c]Significant at .01 level.

monetary control and differences in the reward structure are taken into account.

Table 1 shows the means and standard deviations of all the variables for men and women. The differences in means are generally in favor of men, in some instances appreciably so. The standard deviations are also greater for men, except for weeks worked, and working in a core industry.

Regression results for the total sample with earnings as the dependent variable are shown in Table 2. The first nine independent variables are those generally used in human capital explanations. The results are well in tune with those usually obtained. Of the next two variables, Core Industry and Number of Employees, only one is significant, though both come in with the expected sign. To these standard human capital and institutionalist variables, we have added two measures of work power, Control over Monetary Resources and Number of People Supervised, as well as two sex-related variables, Sex of Supervisor and Sex of Employee. The total equation explains a very respectable 56 percent of the variation in earnings.

As has been demonstrated in numerous other studies, using sex as a dummy variable in a single regression for men and women does not tell the full story, because the reward structure for male and female workers tends to be different. Hence we can learn more about the process that causes the earnings of similar men and women to differ by examining separate regressions.

Table 3 reveals important differences in the reward structure for men and women. Men are rewarded to a substantially greater extent for education. Also, marriage may have a positive effect on their earnings, although the level of significance for men is only .1. Years of Experience has a similar positive effect for both, but there is some evidence that the rise may level off more sharply for women. Finally, working in a Core Industry appears to be valuable for men, though again only at the 10 percent level. This lack of significance for women is not unexpected, for their jobs in the core tend to resemble those in the periphery, and rarely involve upward mobility. Such find-

TABLE 3—DETERMINANTS OF Ln OF EARNINGS OF MEN AND WOMEN

	Men ($N=221$)	Women ($N=195$)
Years of Education	0.0763[c]	0.0401[c]
Years of Experience before Current Employer	0.0220[b]	0.0208[b]
Years of Experience before Current Employer, Squared	−0.0002	−0.0005[a]
Years with Current Employer, Previous Job	0.0127	0.0244
Years with Current Employer, Previous Job, Squared	−0.0001	−0.0006
Years in Current Job	0.0177[c]	0.0185[c]
Weeks Worked	0.0086[c]	0.0098[c]
Hours Worked	0.0137[c]	0.0184[c]
Married	0.1247[a]	0.0747
Core Industry	0.1187[a]	0.0484
Ln Number of Employees of Firm	0.0232[b]	0.0331[b]
Control over Monetary Resources	0.0084[c]	0.0089[b]
Number of People Supervised	−0.0024	0.0119[c]
Sex of Supervisor	0.1429	0.2761[c]
Constant	0.0984	−0.0947
R^2	.48	.44

[a,b,c]See Table 2.

TABLE 4—DETERMINANTS OF MONEY CONTROL FOR MEN AND WOMEN (Tobit Analysis)

	Men ($n=278$)	Women ($n=230$)
Years of Education	6.685[c]	4.161[c]
Years of Experience before Current Employer	.893[a]	2.032[c]
Years of Experience before Current Employer, Squared	−0.020	−0.072[b]
Years in Current Job	1.495[b]	0.783[a]
Years in Current Job, Squared	−0.038[a]	−0.017
Hours Worked	0.483[b]	0.434[c]
Core Industry	−5.279	−9.032[c]
Ln Number of Employees of Firm	−.618	−1.934
Constant	−93.650	−45.264
$-2\times$log likelihood function	1156.99	766.35

[a,b,c]See Table 2.

ings suggest that the core vs. periphery industry distinction, whatever contribution it makes to understanding the situation of male workers, does little to explain differences among women.

Women, on the other hand, gain relatively more from employment in firms with a larger number of workers, possibly because of greater enforcement of equal opportunity laws. Women's earnings increase more with additional hours worked. Net of control over monetary resources, women are rewarded for supervising other workers, but men are not. In our sample, 40 percent of the women have one or more subordinates compared with 49 percent of the men. Also, women, but not men, earn more when they have a male supervisor. Since a worker is far more likely to have a male supervisor in a male occupation, these results may simply show that women earn more in such jobs. Not surprisingly, 94 percent of men and only 57 percent of women have male supervisors.

In spite of these mixed results, the differences in the reward structures make a substantial contribution to the earnings gap. Men actually earn an average of $24,158 as compared to the $18,634 they would earn if they were rewarded according to the women's regression. Women actually earn an average of $12,799 instead of $18,247, which is how much they would earn if they were rewarded according to the men's regression.

The next step is to examine the influence of human capital variables on the attainment of control over monetary resources. The results of a Tobit procedure are seen in Table 4. We find that education has a far stronger positive effect for men than for women as do years on the current job. Most notable, perhaps, is the substantial negative effect for women of being in a core industry, while there is no significant effect among men. This confirms, once again, that the nature of women's work in core industries is different from that of men. Only years of experience before current employer has a larger positive coefficient for women than men.

Somewhat surprisingly, the net result is that both men and women have more control over financial resources than they would have according to the regression for the opposite sex. The difference is, however, substantially greater for male than female workers. The actual figure for men is 20.20, while it would be 17.79 according to the women's regression. The actual figure for women is 16.58, while it would be 15.64 according to the men's regression.

Finally, we compare men's and women's actual earnings with the wages they would receive if they attained monetary control according to the regression for the opposite sex. The result of this procedure is that women with their present qualifications would earn $17,619, or $4,820 more than they presently earn, while men would earn $18,486, or $5,672 less than they presently earn.

II. Conclusions

This study confirms that reward structures for male and female workers are different even when men and women have similar characteristics and are in similar jobs. We also found that women are at a disadvantage in attaining control over work-related resources. The actual earnings gap in our sample is $11,359. It would be reduced to $6,539 if both men and women achieved work power and rewards as men do, or to $5,687 if both achieved work power and rewards as women do. Both attainment process and reward structure make a contribution to accounting for this difference, but the latter is far more important. It is not mainly what women do, but what is done to them that keeps them in an inferior position. Our findings therefore, also suggest that pay according to comparable worth, as well as giving women the opportunity to do comparable work, is likely to be an important component of a program to achieve equity in the labor market.

REFERENCES

Ferber, Marianne A. and Spaeth, Joe L., "Work Characteristics and the Male-Female Earnings Gap," *American Economic Review Proceedings*, May 1984, *74*, 260–64.

Gordon, David, Edwards, Richard and Reich, Michael, *Segmented Work, Divided Workers: The Historical Transformation of Labor in the United States*, Cambridge: Cambridge University Press, 1982.

Hill, Martha S., "Authority at Work: How Men and Women Differ," in Greg J. Duncan and James N. Morgan, eds., *Five Thousand American Families: Patterns of Economic Progress*, Ann Arbor: Institute for Social Research, 1980.

Landes, Elisabeth M., "Sex Differences in Wages and Employment: A Test of the Specific Capital Hypothesis," *Economic Inquiry*, October 1977, *15*, 523–38.

Mincer, Jacob and Ofek Haim, "Interrupted Work Careers: Depreciation and Restoration of Human Capital," *Journal of Human Resources*, Winter 1983, *17*, 3–24.

Spaeth, Joe L., "Job Power and Earnings," *American Sociological Review*, October 1985, *50*, 603–17.

Party Strategies, World Demand, and Unemployment: The Political Economy of Economic Activity in Western Industrial Nations

By JAMES E. ALT*

Whether a change of party control of government in a country with an advanced industrial economy results predictably in a sustained change in its level of aggregate economic activity is continually disputed in political economics. While many issues are involved, the central question is whether constraint ("environment") or discretion ("party control") dominates the explanation of public policy. "Party-induced" changes in public policy demonstrate accountability in democratic government to a political scientist; to an economist they raise the spectre of inefficient outcomes associated with the "political business cycle." In addition, since sustained, party-induced effects on economic policy are inconsistent with the assumptions of many macroeconomic models, economists have also stressed the need for rigorous econometric controls before accepting the existence of such effects.

Douglas Hibbs (1977) has estimated an elegant and rigorous model of party differences in economic policy and activity in Britain and the United States. Briefly, his argument is that 1) blue-collar workers bear the burden of economic contractions (higher unemployment) disproportionately and thus favor government intervention to achieve expansion, and 2) left-wing political parties organize to elicit this blue-collar support and reward their supporters when in office. Doubts about the latter arise from visible counterexamples. Callaghan's British, Mitterand's French, and Carter's U.S. administrations all pursued contractionary policies. All also lost popular support, consistent with the first argument.

However, extending the class-party model to government behavior in other countries presents two complications. First, changes in world economic activity constrain domestic policymakers in smaller countries with open economies from changing the course of economic policy, should they wish to. Indeed, we shall show that unemployment falls under Left governments and rises under Right governments, but predictable effects are systematically visible only relative to the constraint on choice imposed by the world economy.

More important, if political intervention in the economy were costless, then institutions would not matter much. If there are political costs to bear as well as benefits to gain from intervention, then political institutions—here, the organization of electoral, party, and parliamentary systems—structure politicians' incentives, determining whether political intervention in the economy will be present or absent, and, if present, whether interventions are likely to be sustained or transitory. Moreover, these party-induced effects are not "automatic" but are contingent on strategic incentives imposed on politicians by political institutions. In particular, the expected changes in aggregate economic activity appear only when they have been promised. Such party-induced effects are more likely where single-party governments

†*Discussants*: Lewis E. Hill, Texas Tech University; Walt Misiolek, University of Alabama; Jack E. Adams, University of Arkansas-Little Rock.

*Department of Political Science and Center for Political Economy, Washington University, St. Louis, MO 63130. This research was supported by the National Science Foundation grant no. SES-8512037.

form (but are also more likely to be transitory), while under coalitions they are less likely (but if they do exist, are more likely to be sustained). Independent of all these other factors, party-induced effects on economic activity are more likely where parties control parliamentary majorities. Finally, dependence on trade creates a demand for institutions to promote intersectoral bargaining over the distribution of the costs of adjustment, though the success of these institutions in offsetting external shocks is not clearly established by empirical evidence.

I. Party Strategies: Theory

What should we expect economic activity (hereafter, unemployment) to do when we read of a change of party control of government from Right to Left in some small country? The *economic* answer is, independent of the effects of ongoing domestic market processes and world conditions, nothing special. If one knew nothing of politics or the world, one would expect unemployment rates over time to appear autoregressive and fluctuate around some equilibrium level. However, class-party modelers from Kalecki to Hibbs say, expect unemployment to fall instead: parties are policy-oriented, ideological agents of their supporters and will produce this effect *automatically*. My view of competing parties sees them as *strategic* election-oriented, willing to compromise, principals rather than agents, but not autonomous. Hence I say, unemployment will fall if and only if the new government had promised to do something which would lower unemployment, and the promise is kept.

Assume it must have been advantageous for one party or the other to raise economic issues. Why would a promise be kept, once made? It will be easier to answer this in three parts: why kept *at all*; if kept, why *initially*; and if kept initially, why *permanently*? To see why at all, assume voters in an election decide among competing parties by assessing the effects on their interests through both retrospective evaluations of performance and prospective judgements based on party promises, where the value of promises is discounted by the promising party's reputation for reliability or dependability in keeping past promises. Parties, which are organizations of politicians, seek office, make promises to gain it, and have incentives to keep those promises. The desire for credibility—to make more promises and have them believed—and the assumption that reputation for reliability is in the voters' decision calculus ensure that parties find it worth keeping at least some promises, if only to avoid having to run entirely on their records when no one believes their promises.

Why initially? If voters have short memories, as some retrospective voting theorists claim, then it is not obvious that a party which could benefit electorally from keeping a promise would do so right away, for amnesia would dissipate the benefit from doing so before the next election. But early-term actions by new incumbents are important in generating their reputation for reliability, for attention is then focused most clearly on their recent promises, and the excuse that changed circumstances dictate different actions is not credible. Parties could for strategic reasons also not promise publicly to do something they would nevertheless like to do, and then do it anyway, though this seems unlikely.

Moreover, if intervening is costly, then initially is the cheapest time to keep promises, for the "honeymoon" enjoyed by new incumbents means that promises can be kept unopposed which might be strenuously opposed at other times. Consider the honeymoon from the losing side's viewpoint: either they believe the winner's policy is wrong, and that its speedy implementation will be detrimental to the winners in the future, or they fear the winners' policy may be right, and having just lost, need time to reorganize around new people and new issues. Beyond this, since electors wish to see promises kept and both parties wish to keep promises in order to bid for support, and the out-party hopes to be incumbent some time, it pays each party to engage in reciprocal exchange (to a point) by allowing the newly victorious party to keep some promises. However, if the support generated from a continuing intervention is subject to diminishing marginal returns (see my 1985 article), the incentive to maintain an intervention declines. The political costs of intervening include not only those

of creating a coalition, but also the opportunity costs incurred in loss of other programs and the results of agenda changes following government success. Ultimately the largest opportunity cost from maintaining any one intervention is the loss of opportunity to do different things to attract broader support in a future election. Given increasing costs of maintaining an intervention and diminishing returns from doing so, at some point the intervention will stop, or be transitory.

So far, a party forming a government has been treated as a single individual or united team which bids for broad support. In broad coalitions of different parties, incentives are different. In the first place, each party wishes to secure a share in office, to stay in office by securing the support of its own bloc. Bargaining over policy changes which might lead to transitory interventions in single-party governments comes before government formation. There may be no policy agreeable to all coalition partners, or if there is a bargain, it may be the only agreeable bargain and thus be sustained. If one party in a coalition wishes to change the course of policy, it may be frustrated by other partners who seek to deny it advantage. They will threaten to break up the coalition, in which case either the original policy survives, or the threat is carried out and the government breaks up. Either way the original intervention has been sustained throughout the incumbency.

So the strategic view says that, because of prospective voting, if intervention has been promised, then it is likely to appear. The combination of retrospective elements in voting plus costly intervention means that in single-party governments, the intervention is likelier to be carried out but less likely to be sustained; in coalitions it is less likely to be carried out but, if it is, more likely to be sustained. If U-turns in policy reflect only economic constraints, these differences would not appear according to party systems. If effects are automatic, they will appear regardless of the presence of explicit promises. (Naturally, other things equal, effects are always more likely where governments have parliamentary majorities than minorities.) Finding variations corresponding to promise making and party systems thus supports assuming the presence of retrospective and prospective elements in voting, costly intervention and thus the importance of institutions in economic policy, as well as the idea of parties as strategic actors.

II. World Demand: Theory

Increasing liberalization of trade since World War II has made the economies of many advanced industrial economies more "open," more dependent on trade. Among countries with open economies, those that are "large" pursue and preserve national autonomy as a goal of policy while those that are "small" (like all price takers) adjust. Most industrial countries are small and have open economies: for them, the state of the world economy is a formidable constraint and a decline in levels of world economic activity would normally increase domestic unemployment. This increase could be offset by reducing wage costs, unlikely if openness produces strong labor movements (David Cameron, 1978), or profits, unlikely if profits are sustainable at lower production levels, or by trading productivity losses for increased employment, as done in crises in Austria through large state sector adjustments (Ewald Nowotny, 1983), reallocating production with attendant conversion costs, or subsidizing exports, in spite of trade wars and sectoral competition. Otherwise, the response of domestic unemployment to an international shock should be proportional to the share of international trade in the domestic economy.

However, in this way a country's dependence on trade (hereafter: openness) magnifies the demand among for workers for Left government and institutions of broad collective bargaining. This comes about in several ways. The impact of external shocks on domestic employment creates the demand for insurance which can be exploited by Left parties. But external effects affect sectors within classes differently (Rachel McCulloch, 1983). Sustaining low unemployment against increases in world levels thus requires suppressing sectoral conflict among workers, so that public action does not produce selective benefits to competing groups. Exposure to international competition induces industrial concentration, resulting in centralization and

a high degree of worker organization necessary to demand reduced unemployment and Left government to supply it (Cameron, 1978). A labor federation sufficiently all-embracing and involved in collective bargaining to act as a monopolist could secure and enforce wage and employment bargains across the board to prevent the loss of aggregate worker benefits through conflict among worker organizations. Philip Schmitter (1981) describes "corporatist" associational monopoly characteristics of labor federations like broad membership, absence of factions, and dominance of one labor organization. Cameron (1984) correlates these characteristics, scope of collective bargaining, and existence of workers councils with economic openness and long-run levels of inflation, and implies that Left government and some of these evolved institutional forms also explain different unemployment rates among countries. While the general correlation between openness, corporatist, and in some circumstances Left government is clear, evidence of institutional effects on unemployment levels is less decisive.

III. Party Strategies: Results

Measurement of the effects described above requires dates of changes, majority status, and party composition of governments, a measure of the salience of the economy as a political issue, and domestic and world unemployment rates, all of which were collected for fourteen major industrial countries for 1960–83. Details are given in my earlier article. Dynamic models for unemployment were estimated (an ARIMA model for domestic unemployment with a transfer function for world demand and intervention terms for party control changes) initially treating the effects of changes separately as sustained or transitory, with lower error sum of squares indicating which model was to be preferred in each country. These final models produce significant effects for about one change of government in four, not very impressive though unlikely to arise entirely by chance. However, while the economy was a salient issue in only half the changes of government, over 80 percent of the significant party-induced effects occur when the economy was an issue, consistent with the argument that party-induced effects result from promise keeping. That promises are only made half the time, and party-induced effects rarely occur when not promised, indicates that both the promise and its keeping are not automatic but rather contingent on the strategic advantage of competing parties.

In the five countries where broad coalitions predominantly form governments, the model of sustained effects fit better in four cases; in the eight where single-party or dominant-partner governments form, the model of transitory effects fit better in six. Where the economy was an issue and the "expected" model fit better (i.e., omitting Austria, Finland, and Sweden, and all other cases where economic change was not promised), then significant transitory party-induced effects appear 85 percent of the time but significant sustained effects appear only 40 percent of the time. Effects of any sort appear less under minority governments. So early-term effects on unemployment appear to reflect the strategic keeping of strategically made promises, but taking office in a minority administration or broad coalition is likely to frustrate keeping these promises, though promised effects, if carried out, are more likely to be sustained in coalitions.

IV. Adjustment to World Demand: Results

Apart from these effects of partisan advantage, there is a further question of whether institutional differences among countries explain their different abilities to adjust to external shocks. Tests for econometric causality make it clear that changes in world unemployment are led by recent prior changes in American unemployment (though the induced changes in world activity then independently feed back onto the American economy) with possible leading effects on world unemployment from France and Germany as well. There are significant delays in the importing of shocks from the world into the Scandinavian economies, but otherwise shocks appear to be transmitted among countries quickly, on the order of one or two months.

The cumulative impact of a unit shock in world demand on domestic unemployment

relative to average unemployment in that country measures the extent to which domestic unemployment responds to the world economy (data are given in my earlier article, Table 6). This responsiveness correlates closely with Ray Fair's (1982) estimates of the response of exports to a shock in American real *GNP*, though less well with his estimates of the response of other countries' *GNP* to an American-induced real shock. Openness explains variations in unemployment responsiveness, which increases by about 0.6 percentage point for each extra percentage point of exports in *GDP*, though the increase is less than proportional in the more open economies. However, independent of openness, no political-institutional factors (using data from Cameron, 1984) significantly reduce this elasticity: scope of collective bargaining, Left party control, and the interaction of party control with associational monopoly have, as outlined above, negative effects but are not statistically significant singly or jointly. Average unemployment 1960–83 increases with openness, implying that the cumulative effects of world recession explain individual countries' experience of unemployment. Independent of the positive effect of openness, the existence of workers' councils and share of labor force unionized both reduce average unemployment significantly but party control has no independent significant effect. This suggests that the organization of institutions like party systems and labor confederations shape not only incentives for competing parties, but also possibilities for enforcing broadly bargained macroeconomic outcomes.

REFERENCES

Alt, James E., "Political Parties, World Demand, and Unemployment," *American Political Science Review*, December 1985, *79*, 1016–140.

Cameron, David, "The Expansion of the Public Economy: A Comparative Analysis," *American Political Science Review*, December 1978, *72*, 1243–61.

______, "Social Democracy, Corporatism, Labour Quiescence and the Representation of Economic Interest in Advanced Capitalist Society," in John Goldthorpe, ed., *Order and Conflict in Contemporary Capitalism*, Oxford: Oxford University Press, 1984.

Fair, Ray, "Estimated Output, Price, Interest Rate, and Exchange Rate Linkages Among Countries," *Journal of Political Economy*, June 1982, *90*, 507–35.

Hibbs, Douglas, "Political Parties and Macroeconomic Policy," *American Political Science Review*, December 1977, *71*, 1467–87.

McCulloch, Rachel, "Unexpected Real Consequences of Floating Exchange Rates," in *Essays in International Finance*, No. 153, Princeton, 1983.

Nowotny, Ewald, "Nationalized Industries as an Instrument of Stabilization Policy," *Annals of Public and Cooperative Economy*, March 1983, *53*, 41–57.

Schmitter, Philip, "Interest Intermediation and Regime Governability in Contemporary Western Europe and North America," in Suzanne Berger, ed., *Organizing Interests in Western Europe*, Cambridge, New York: Cambridge University Press, 1981.

What Can Economics Learn from Political Science, and Vice Versa?

By K. ALEC CHRYSTAL AND DAVID A. PEEL*

The concerns of economics and political science intersect. Both have as a major object of enquiry the making of economic policy. Indeed, both have a reasonable claim to primacy in this field. Economic policy is made by incumbent politicians in the context of political institutions. The analysis of the impact of such policies as emerge is, on the whole, the job of the economist. Economic analysis influences the ideas of politicians, but it is the political scientist whose job it is to illuminate the political decision-making process itself. While both have this interest in economic policy in common, their methods and concerns differ considerably. What can each learn from the other?

In Section I, we argue that so-called "Politico-Economic" (*P-E*) models represented a false start in the attempt to define the common ground. Section II gives an example of how economists' way of thinking can be a fruitful source of testable hypotheses for political science, and in Section III, we argue that economists need to pay much more attention to institutional arrangements which are the "bread and butter" of political science.

I. Politico-Economic Models: A False Dawn

In the 1970's it seemed as if a new literature was emerging which was common to economics and political science. This started off as a model of the political business cycle (William Nordhaus, 1975) but was subsequently labelled *P-E* (Bruno Frey and Friedrich Schneider, 1978). The *P-E* models had two key features. First, government macro policies were presumed to depend upon political factors such as the proximity of an election or the popularity of the party in power. Second, the popularity of the government was supposed to be stably related to economic indicators such as inflation or unemployment.

Unfortunately, neither of these relationships proved to be robust either from a theoretical or an empirical perspective. Popularity functions have proved to be extremely unstable while political variables which systematically improve the performance of macro forecasting models, *ex ante*, are elusive. Indeed, the gulf between the *P-E* perspective and mainstream macroeconomics has widened considerably in the past decade. This is because the most popular versions of the *P-E* approach rely upon the government being able systematically to fool the electorate about the true state of the economy at election time. With well-informed actors no such electoral advantage can be achieved.

Theoretical inadequacies can often be forgiven if an approach is empirically robust. However, the *P-E* models fail most badly on this score. Any *P-E* model would have predicted a huge defeat, rather than a resounding victory, for Mrs. Thatcher in the 1983 election. Unemployment had more than doubled in her 1979–83 term of office. Of course, there were special factors at work. But every election is special—that is the point.

II. A Rational Expectations Model of Popularity

The rational expectations hypothesis has been widely applied in economics. It is clearly relevant to the formation of expectations of the performance of political parties. Suppose actors support the party that offers them the greatest present value of expected future benefits. "Benefits" could in principle be much broader than just income and could

*Professor, University of Sheffield, England, S10 3EY, and Professor, University of Wales, Aberystwyth, SY23 3DB, respectively. Computing assistance of Dot Jones and Natalie Woodcock is gratefully acknowledged.

include anything that affects subjective feelings of well being. The percentage of the electorate supporting party X at time t can be written

$$PX_t = \alpha PVX_t + U_t, \quad (1)$$

where PX is the percentage support for party X, PVX is the *net* present value of having party X in power as opposed to any other party, α is a positive constant, and U is a random measurement error. This may seem fairly vacuous since the term PVX, being based upon expectations of *any* potentially relevant factor, is virtually impossible to measure. However, the rational expectations assumptions does give testable content, since only new information will lead to changes in PVX. This "news" is by definition unforecastable and is, therefore, white noise. Thus

$$PVX_t - PVX_{t-1} = \varepsilon, \quad (2)$$

where ε is a white noise error.

Differencing (1) and substituting (2) leads to

$$\Delta PX_t = \alpha\varepsilon + U_t - U_{t-1}, \quad (3)$$

where Δ is the difference operator. The composite error term in (3) can be written as a moving average process of order one, with a negative parameter whose size depends on the relative magnitudes of α and the variances of ε and U (see Andrew Harvey, 1981, p. 431). Consequently popularity is hypothesized to follow an ARIMA (0,1,1) process. An additional test of this hypothesis is provided by the fact that the change in popularity should be orthogonal to any information that was available last period. So lagged economic or political variables should not have any impact, since, they would already have influenced last periods popularity.

We have performed tests on monthly popularity data for six countries (Canada, U.K., New Zealand, France, West Germany, and Norway). This involved a total of 20 political parties. In addition we used monthly data on the popularity of the president of the United States. These data were subjected to two sets of tests. The first was the estimation of the ARIMA process and the second was the orthogonality test. Data periods varied in length from 1947(2)–1985(3) for the U.K. to 1977(6)–1983(3) for Norway. In the U.S. case, a shift dummy was included for the January change over of president.

The first test resulted in a remarkable degree of consistency. (Results are available on request.) In *all* cases, popularity was found to be describable by either an ARIMA (0,1,1) or (0,1,0) process. The Box-Pierce statistic at 10 and 20 lags was consistent with the hypothesis of no higher-order autocorrelation. In this respect, support for the hypothesis could not be stronger.

The orthogonality test involved the inclusion of lagged inflation and unemployment as explanatory variables. In addition we used an interelection cycle and both rising and falling trends. These latter variables have been claimed to fit well in previous studies. Except for the United States, a slope dummy was included in this test to allow for a differential effect when a party was in power. For the United States, U.K., Canada, and New Zealand, all of these additional variables were insignificant, thus confirming the hypothesis of orthogonality. Some rejections were found in France for the Socialists, in West Germany for the SDP and Free Democrats, and in Norway for the Conservatives, the Christian Peoples Party and the Centre Party. However, the rejections occur in only 14 of 95 test coefficients. This is little more than could arise by pure chance. When combined with the parsimonious representation of popularity data by the ARIMA process, this appears to be a promising start for a new approach to the interpretation of popularity data. It certainly provides one explanation of why the previous literature has made such heavy weather of the search for a stable popularity function.

III. Political Science and the Failure of Economists

Economists are very good at a priori reasoning. Testable hypotheses can be generated by deductions from optimising assumptions. This is especially easy where the actors can be encapsulated in a perfect market. In this

way the model need not be tarnished with the realities of a pecular institutional structure. For the economist, once special circumstances have to be allowed for this is "*ad hoc*" and, therefore, inferior from a methodological perspective. This aversion for the understanding of institutional context leads economists to make gross errors of analysis. These are particularly serious when policy advice is exported to another country. Often this is done in ignorance of critical institutional differences—industrial structure, union power, monetary institutions, exchange rate arrangements, degree of openness, etc. Those who import models, developed for other contexts, without question are more guilty than most. Here, there is no excuse for not understanding the institutional structure.

One example is available from the "money surprise" literature. Following the pioneering work of Robert Barro (1977) in the United States, a number of authors applied his approach to other countries. The key idea is that only unexpected money growth will trigger real business cycles since anticipated money will be reflected in price setting and not output. The hypothesis requires the money stock to be an exogenously determined variable. It is reasonable to test it in the U.S. context (though even in the United States this may be questionable if the authorities are pegging interest rates), however, in a country like the U.K., the direct application of the Barro model to much of the postwar period is totally inappropriate. Until 1972, the U.K. had a fixed exchange rate. Consequently the money stock cannot be an exogenous variable. Hence causation is reversed.

Nonetheless, several studies apply the money surprise idea unquestioningly as if money were exogenous (for example, Cliff Attfield et al., 1981). No attempt is made to justify this assumption despite the theoretical and empirical support for reverse causation for the fixed exchange rate period in the U.K. (David Williams et al., 1976). Even post-1972, it is questionable in the U.K. context whether the money stock was exogenous. This is especially true of broader monetary aggregates such as $M3$ which has commonly been used in the money surprise studies, due to the U.K. authorities' predilection for controlling interest rates rather than the money stock.

TABLE 1—CAUSALITY TESTS BETWEEN MONEY, INTEREST, AND ACTIVITY FOR THE U.K.

(1)	(2)	Col. 1 Caused by Col. 2	Col. 1 Causes Col. 2
$M0$	*IND. PROD*	1.96[a]	2.58[b]
$M0$	*PRICES*	0.85	1.74[a]
$M0$	*UNEMPL*	1.34	1.71[a]
$M0$	*INTEREST*	1.74[a]	1.12
$M1$	*IND. PROD*	2.45[b]	1.02
$M1$	*PRICES*	0.41	2.36[b]
$M1$	*UNEMPL*	2.53[b]	1.35
$M1$	*INTEREST*	2.27[b]	1.1
$M3$	*IND. PROD*	1.9[a]	0.5
$M3$	*PRICES*	2.11[b]	0.92
$M3$	*UNEMPL*	2.51[b]	1.05
$M3$	*INTEREST*	1.57	1.25
INTEREST	*IND. PROD*	0.99	0.51
INTEREST	*PRICES*	0.64	1.46
INTEREST	*UNEMPL*	0.77	2.17[b]

Notes: Data were monthly unadjusted 1973:4–1985:2. Causality tests include 18 lags of dependent and independent variables. $M0$ is the monetary base; INTEREST is the 90-day Treasury bill rate.

[a,b]The critical F-statistics with 18 and 106 degrees of freedom is 1.7 at the .05 level and 2.1 at the .01 level. Footnote keys indicate the rejection of the hypothesis of no causality at these respective probability levels.

Table 1 reports tests of Granger causality between three monetary aggregates ($M0, M1, M3$) and industrial production, prices, unemployment and the interest rate. This represents part of the test for what Robert Engle et al. (1983) call "strong exogeneity." On this basis we would have to reject the hypothesis of exogeneity for all three of the monetary aggregates. The interest rate appears to be a much stronger candidate for exogeneity. It is caused by no other variable and yet is causal of both $M0$ and $M1$. Of course, there are a number of interpretations of the interest rate evidence, for instance, that it is policy determined either in London or Washington. However, the crucial point is that no political scientist would blindly apply a model developed in one country to another without a thorough

investigation of the context in which the application was to be made. Indeed, the political scientist would often only regard the transference as worth making if the situation were sufficiently *different* that the implications of the differences could be tested. Economists tend to be arrogant because their techniques are more advanced than those of other social sciences. They should be humble that they still have so much to learn.

IV. Summary

Economists have made great advances by the application of abstract mathematical models. The "best" of these are based upon the optimizing behavior of actors. The development of tractable models has required the disregard of many institutional details. This has benefits—but there are costs. Theoretical advances will only be of wider social benefit if they are relevant in the real-world institutional context. Political science has the opposite problem. It has been so concerned to get the detail right that the idea of statistical regularity and hypothesis testing, not to mention model building, have been treated with the greatest suspicion. Both have much to teach the other.

REFERENCES

Attfield, Clifford, L. F., Demery, David and Duck, Nigel W., "A Quarterly Model of Unanticipated Monetary Growth, Output and the Price Level in the U.K.: 1963–1978," *Journal of Monetary Economics*, November 1981, *8*, 331–50.

Barro, Robert J. "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

Engle, Robert, Hendry, David F., and Richard, Jean-Francois, "Exogeneity," *Econometrica*, March 1983, *51*, 277–304.

Frey, Bruno S. and Schneider, Friedrich, "A Politico-Economic Model of the United Kingdom," *Economic Journal*, June 1978, *88*, 242–53.

Harvey, Andrew C., *Time Series Models*, Deddington: Phillip Allan, 1981.

Nordhaus, William D., "The Political Business Cycle" *Review of Economic Studies*, April 1975, *42*, 169–90.

Williams, David, Goodhart, Charles A. E. and Gowland, David H., "Money, Income and Causality: The U.K. Experience" *American Economic Review*, June 1976, *66*, 417–23.

Political Parties and Macroeconomic Policies and Outcomes in the United States

By DOUGLAS A. HIBBS, JR.*

Containing inflation at politically acceptable rates of unemployment has been the most important macroeconomic policy problem confronting American political authorities for several decades. In a democratic political system, macroeconomic policies are rarely, if ever, motivated by apolitical Golden Rule norms; rather they are conditioned, at times, decisively, by political forces. In this paper I analyze one important source of political influence on postwar macroeconomic policies and outcomes in the United States: the differing economic goals of the political parties.

I. Party Constituencies, Voter Priorities, and Party Goals

In American national politics, the core constituency of the Democratic party consists of the down-scale classes, who primarily hold human capital and bear a disproportionate share of the economic and broader social costs of extra unemployment. Up-scale groups form the core constituency of the Republican party; they hold financial capital and absorb the greatest losses from extra inflation. For this reason Democratic voters generally express greater aversion to unemployment and less aversion to inflation than Republican voters (see my forthcoming book, chs. 3–6, and the studies cited).

Differences in the economic interests and revealed preferences of the parties' core constituencies are reflected in the pattern of policies and outcomes observed under Democratic and Republican presidential administrations. Democratic administrations have been more likely than Republican ones to pursue expansionary policies yielding lower unemployment and extra growth, but running the risk of higher inflation. Republican administrations typically weight the problem of inflation more heavily. Consequently, they have more readily and more vigorously pursued disinflationary policies, and in general they have been more cautious about stimulating aggregate demand and employment.

In what follows I use stylized theoretical models to obtain econometric estimates of the party cleavages just described. The analysis here is brief and confined to unemployment outcomes and monetary policy under the parties. Elsewhere I use the same framework to analyze real output and fiscal policy and develop the story in greater detail (see my book, chs. 8–10; for an earlier analysis, see my 1977 article).

II. Unemployment under the Parties

The impact of the Democratic and Republican parties' different economic goals on the time path of unemployment may be evaluated by the following stylized model. Both have different unemployment targets, U^T, that are constrained by, and therefore tend to vary, around a "normal" or benchmark unemployment rate, U^N. The unemployment target prevailing during Democratic presidential administrations is lower than the corresponding target during Republican administrations, which is represented by the (quarterly) U^T equation

$$U_t^T = \beta_0 + U_t^N + \beta_1 Dem_{t-1}, \tag{1}$$

where *Dem* is a binary variable equal to $+1$ during Democratic administrations and 0 during Republican administrations; U^N is set equal to Robert Gordon's (1984) calculation of the natural unemployment rate; and the party hypothesis requires $\beta_1 > 0$. The *Dem* term appears with a one-quarter lag because the unemployment target reflected in

*Harvard University and Department of Political Science, Göteborg University, 40221 Göteborg, Sweden.

current policies is based on the party in power during the previous period.

Given behavioral lags in policy formulation, institutional lags in policy implementation and, most important, structural lags in the response of the economy to policy actions, presidential administrations cannot achieve their economic objectives immediately. Therefore, administrations are able to adjust the observed unemployment rate, U, to their preferred rate, U^T, only partially each period. The adjustment mechanism is

$$(2)\quad U_t - U_{t-1} = \phi_1(U_t^T - U_{t-1}) + \phi_2(U_{t-1} - U_{t-2}) + \beta_2\, Shock_{t-1} + e_t,$$

where $0 < \phi_1, \phi_2 < 1$, and e is a well-behaved disturbance.

Hence, policy-induced changes in unemployment from one quarter to the next are capable of closing only a fraction (ϕ_1) of the gap between the current target and the actual unemployment rate observed for the previous period. The remaining right-hand side terms in (1) add a bit more realism to the adjustment model. Additional structural inertia in the time path of unemployment is accommodated by the lagged rate of change term. Fluctuations in unemployment due to shocks exogenous to the domestic political economy, notably the energy price hikes of 1973–74 and 1979–80, are represented by the variable *Shock*, which is equal to the percentage point rise in the price of imported oil, weighted by the net share of oil imports in *GNP*.

Substituting (1) into (2) and solving for U_t yields a nonlinear second-order estimating equation for the time path of unemployment:

$$(3)\quad U_t = \phi_1\beta_0 + (1-\phi_1+\phi_2)U_{t-1} - \phi_2 U_{t-2} + \phi_1 U_t^N + \phi_1\beta_1 Dem_{t-1} + \beta_2 Shock_{t-1} + e_t.$$

Estimation results for equation (3) in Table 1 show that the coefficients of all terms have the anticipated signs and satisfy conventional statistical significance levels. The rate of adjustment of unemployment to party targets (ϕ_1) is on the order of 7.5 percent per period, but it is perturbed by structural inertia from the lagged rate of change (ϕ_2). The estimate of the β_2 parameter indicates that each unit rise in the *Shock* term initially increased the unemployment rate by about 0.3 points, after a one-period lag. However, the *Shock* variable reached peak values of 1.6 in 1974:3 and 1.5 in 1980:1, and dynamic calculations suggest that the energy price shocks ultimately raised unemployment about 2 percentage points above U^N in 1975 and again in 1980.

TABLE 1—ESTIMATES OF THE PARTY CLEAVAGE MODELS FOR UNEMPLOYMENT AND MONETARY POLICY, 1953:I TO 1983:II[a]

	Unemployment Rate (U) Equation (3)	$M1$ Growth Rate (m) Equation (6)
α		1.31 (0.55)
β_0	0.843 (0.550)	
δ		−0.875 (0.270)
φ_1	0.076 (0.020)	0.638 (0.085)
φ_2	0.561 (0.072)	
β_1	−2.11 (0.94)	−1.97 (1.04)
β_2	0.313 (0.090)	
$c_1 + c_2$		0.432 (0.148)
Adj. R^2	.963	.361
SE of Regression	0.322	2.73

[a]Estimated standard errors are shown in parentheses.

For my purposes here, the most important parameter is β_1 (*Dem*), which estimates the magnitude of the cross-party difference in unemployment targets and, therefore, the impact of sustained changes in party control of the presidency on unemployment outcomes, after all lags of adjustment. In the unemployment equation, the estimate of β_1 is approximately -2.0, which implies that after adjustment lags, unemployment tends to be about 2 percentage points lower under the typical Democratic administration than under the typical Republican one. Alternatively, this estimate is theoretically consistent

with (and empirically indistinguishable from) the idea that the parties have different ideas about the "normal" or "natural" unemployment level, and this prompts them to pursue different unemployment targets.

The estimated values of ϕ_1 and ϕ_2, which define the rate at which actual outcomes are adjusted to party targets, indicate that the unemployment equation is stable, and that convergence to equilibrium values is nearly complete after 16 quarters, or one presidential term. Hence, if we take all parameter estimates at face (point) values, and assume that U^N now lies between 6 and 6.5 percent, the results indicate that the typical contemporary Democratic administration will aim for (and, after one term, achieve) an unemployment rate just above 5 percent, as compared to a target rate just above 7 percent for the typical Republican administration.

III. Monetary Policy under the Parties

Thus far I have tried to show that the partisan stripe of presidential administrations has significantly affected unemployment outcomes in the postwar American economy. However, policy authorities do not control directly macroeconomic outcomes; they control macroeconomic policies. Monetary policy is easier to maneuver in the short run than fiscal policy, and it is more decisive: the impact of fiscal initiatives are largely dissipated unless monetary growth rates are accommodating. Although the Federal Reserve has considerable formal autonomy under American institutional arrangements, it's insulation from political direction is largely illusory. Numerous studies have concluded that the administration's macroeconomic goals are what drive Federal Reserve policy behavior, as contrasted to Federal Reserve policy rhetoric (the evidence is summarized in my book, ch. 1). This view underlies the theoretical framework presented below.

Letting m denote the (annualized, quarter-on-quarter) percentage growth rate of the $M1$ money supply, we have the target equation

$$(4) \quad m_t^T = \alpha_0 + \delta\left(U_t^T - U_{t-1}\right) + cp_t^*,$$

where p^* denotes the "ongoing" inflation rate (of the implicit GNP deflator), and the party cleavage hypothesis requires $\delta < 0$. Hence, the money supply growth rate target of an administration is proportional to the gap between its unemployment target (U^T) and the actual unemployment rate (U) observed for the previous quarter. When unemployment is above the target, administrations seek to close the gap by pushing the Fed to raise the money supply growth rate (relative to the inflation rate). The reverse is true if the gap between U and U^T goes the other way. The monetary growth target is also conditioned by the ongoing inflation rate, p^*, since movements in the real money supply are what move unemployment and real output. If administrations (and, more directly, the Federal Reserve) are indifferent to inflation, c (the coefficient of p^*) should be in the vicinity of 1.0. If c is less than 1.0, monetary policy goals do not fully accommodate inflationary trends, which implies that unemployment goals may be adjusted upward in order to fight inflation.

The adjustment-to-target equation for the money supply growth rate is

$$(5) \quad m_t - m_{t-1} = \phi_1\left(m_t^T - m_{t-1}\right) + e_t,$$

where $0 < \phi_1 < 1.0$. This adjustment function includes no lagged (inertia) terms because, in principle, the money supply growth rate can be brought into line with party targets quickly. Substituting (4) into (5) (and recalling that U^T is given by equation 1) and solving for m_t, yields the nonlinear estimating equation

$$(6) \quad m_t = \alpha + (1-\phi_1)m_{t-1} + \phi_1\delta\left(U_t^N + \beta_1 Dem_t - U_{t-1}\right) + \phi_1 cp_t^* + e_t,$$

where $\alpha = \phi_1(\delta\beta_0 + \alpha_0)$, and other terms are as defined previously.

The results for equation (6) are reported in the second column of Table 1. The estimate for δ is about 0.9. After lags of adjustment, nominal $M1$ therefore tends to grow nearly point-for-point with deviations of unemploy-

ment from the party targets. The estimate for the speed of adjustment (ϕ_1) of m to m^T is 0.64 (that is, 64 percent per quarter). As would be expected, this is much larger than the corresponding rate of adjustment of U to U^T.

From the point of view of a political analysis of the American economy, the most important parameter, as before, is β_1, which distinguishes the monetary targets and, hence, the unemployment targets of Democratic and Republican administrations. The β_1 point estimate is -2.0; almost identical to the estimate in the unemployment equation. Since the estimates of β_1 in the m and U equations were obtained independently, their near equivalence is strong evidence in favor of the party cleavage model.

However, monetary policy has also responded to the "ongoing" inflation rate (which is measured by annualized, quarter-on-quarter inflation rates lagged one and two periods). The coefficient of p^* (the sum $c_1 + c_2$) is significantly less than 1.0. Consequently, real money supply growth rates, which, as I noted earlier, are what affect unemployment (and real output) movements, have been constrained by inflation rates. (Regressions not reported here reveal that the inflation sensitivity of monetary policy does not vary with the party controlling the White House.) In fact, the estimation results imply that if inflation is high enough relative to the gap between actual and target rates of unemployment, the real money supply turns negative which, after a lag, would tend to yield rising unemployment (and lower future inflation). But unemployment, inflation and (lagged) monetary growth rates are mutually endogenous, and this complicates formal analysis of the model's long-run properties.

IV. Discussion

It is not possible to say from the econometric evidence whether the inflation sensitivity of monetary policy represents relaxation of unemployment goals by presidential administrations or "independent" anti-inflation activity by the Federal Reserve. However, since the unemployment model is essentially a reduced-form equation in which the policy variables have been solved out, we may gain more information about the impact of inflation on operative unemployment goals by entering lagged inflation variables directly into the U^T function, which amounts to estimating equation (3) after adding terms of the form: $\phi_1 \Sigma c_j p_{t-j}$, $j > 0$.

Taking this approach, which circumvents some of the feedback problems noted above, produces estimates of Σc_j that are uniformly positive and range in magnitude between 0.08 and 0.15, but that invariably are not significant at the usual test levels. One might conjecture from such results that there is some tendency for the parties' unemployment goals to be raised on the order of one-tenth of a percentage point per point of sustained inflation. Despite the absence of firm statistical evidence, a conclusion along these lines is surely sensible. After all, even the Democratic party's core constituency dislikes inflation, although less so relative to unemployment than Republican supporters.

Moreover, if the normal or benchmark unemployment variable used in the unemployment model, U^N, is a true natural rate, then unemployment held below U^N indefinitely would yield hyperinflation (which, assuming an unbounded Phillips curve, is what the estimates for equation (3) imply would be the case under a prolonged run of Democratic administrations), and unemployment held above U^N indefinitely would yield hyperdeflation (which is the steady-state result for equation (3) under the Republicans). Since neither party has held the presidency for more than two terms in succession in the postwar (estimation) period, such unrealistic long-run results do not weaken the validity of the model for party regimes of the duration actually observed. Nonetheless, it must be remembered that party cleavages are over the relative emphasis give to macroeconomic problems. A typical administration of either party will ultimately respond to a dominant macroeconomic problem, even if it requires distasteful actions inconsistent with usual priorities.

Yet, postwar electoral history shows that the Democrats have been more likely than the Republicans to get into difficulty with the voters by pursuing overly ambitious unem-

ployment goals creating extra inflation. The Republicans, on the other hand, have more frequently suffered electoral setbacks because of their enthusiasm for disinflationary bouts of economic slack. In an era when elections have increasingly turned on economic performance, such "overshooting" may help explain why neither party has managed since World War II to hold the presidency for more than two consecutive terms.

REFERENCES

Gordon, Robert J., *Macroeconomics*, 3rd ed., Boston: Little Brown, 1984, Appendix B 2.

Hibbs, Douglas A. Jr., *The American Political Economy: Macroeconomics and Electoral Politics in the United States* Cambridge: Harvard University Press, forthcoming.

______, "Political Parties and Macroeconomic Policy," *American Political Science Review*, December 1977, *71*, 1467–87.

Party Differences in Macroeconomic Policies and Outcomes

By HENRY W. CHAPPELL, JR. AND WILLIAM R. KEECH*

Although the nature of the differences between parties in democratic electoral politics is an enduring question in political science, surprisingly little is understood about the subject. But substantial progress has been made in recent years, most notably in understanding party differences in macroeconomic policies and outcomes. The first breakthrough was Douglas Hibbs's (1977) analysis of party-related differences in the unemployment rate. In his time-series analysis for the United States, Hibbs modeled the path of unemployment as an autoregressive-moving average process subject to a dummy variable intervention term indicating party of the president. His analysis indicated that Democratic administrations were associated with lower unemployment than Republicans by 2.36 points after eight years in office, and even larger differences in "long-run equilibrium." A subsequent article by Nathaniel Beck (1982) addressed the same issue, and found the party differences less sharp when administration-specific policy differences are considered.

The techniques employed by Hibbs and Beck focus directly on an outcome (unemployment), rather than on the policy instruments that are presumably responsible for altering outcomes. This approach can be misleading when there are long lags between implementation of policies and ultimate effects, or when shocks occasionally intrude upon the regular connections between instruments and outcomes. Macroeconomic theories can provide information about constraints linking macroeconomic variables, but Hibbs and Beck fail to incorporate theoretical constraints. Such constraints could help determine what kinds of outcomes are feasible and sustainable, and to what extent outcomes are induced by policies as opposed to shocks. Our purpose here is to consider how one might go about estimating party differences in a framework that takes advantage of some insights offered by macroeconomic theories, and to report some preliminary results. (A more complete description of the analysis is provided in our working paper, available upon request.)

I. Traditional Policy Analysis: An Application

One way to assess party differences, while accounting for some of the previously noted difficulties, is suggested by the traditional approach to policy analysis in economics. Systematic policy differences under Democratic and Republican presidents should first be detectable in the paths of policy instruments. Empirical estimates of appropriately specified policy reaction functions should reveal evidence of any party-related differences. To link these policies to resulting outcomes, one can construct a macroeconometric model. Theoretical constraints can be embedded into the structure of such a model, and patterns of lagged response to shifts in policy instruments can be determined in the task of estimation. To assess the impact of party policies on outcomes, the policy reaction functions can be added to the econometric model, and then simulations undertaken under alternative assumptions regarding party in power. Such experiments conceptually control for the impact of stochastic shocks and appropriately assess responsibility for policies which have long-enduring lagged consequences.

In a preliminary application of this approach, we have made use of a version of the St. Louis model that has been augmented by an empirically estimated monetary policy re-

*Department of Economics, University of South Carolina, Columbia, SC 29208, and Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, respectively. We thank John Addison, Michael Bordo, and John McDermott for helpful comments. John Tatom provided the St. Louis model and associated data. This research was supported by National Science Foundation grants no. SES 8420122 and SES 8420709.

action function with party differences.[1] The reaction function specifies that the rate of growth of the money supply ($M1$) is a function of its own lagged value, and of once lagged values of the difference between the unemployment rate and its high-employment level, the inflation rate, and the ratio of the full-employment budget deficit or surplus to high employment *GNP*. In addition, party differences in the intercept are permitted through the introduction of a dummy variable indicating the party of the incumbent president. Estimation of the reaction function using quarterly data from 1953:I to 1984:IV provided plausibly signed and statistically significant (at the .05 level, two-tailed tests) coefficients for all explanatory variables. Most importantly for our purposes, there was evidence of significantly faster money growth on average under Democrats.[2]

We have simulated the augmented model for the period 1953:I to 1984:IV assuming first that Republicans and then that Democrats were in office for the entire period. We then compared unemployment paths under the two scenarios. Our results indicate that the initial impact of replacing Republicans with Democrats was to lower the unemployment rate—after four years the unemployment rate was just over a point lower under the Democrats. Beyond the four-year point, the difference in the unemployment rates gradually declined, although a slightly lower rate (by .2 points) for Democrats actually persisted for the remainder of the simulation.[3] Inflation was permanently higher under the Democrats by about 2.5 percentage points.

[1]We ignore possible party-related impacts of fiscal policies. Impacts of fiscal policies are empirically weak in the St. Louis model; moreover our reaction function estimates did not provide robust evidence of party differences in fiscal variables.

[2]Robustness of the party difference effect is a concern. Using alternative economic indicators as regressors or different periods for estimation, the party effect was sometimes insignificant.

[3]The St. Louis model does not impose a natural rate property for the long run, a priori. However, the simulation results are at least approximately consistent with that hypothesis.

The four-year impacts of party on unemployment are quite similar in magnitude to those reported by Hibbs, but beyond four years the implications differ substantially. The Hibbs approach extrapolates short-run effects to the long run, whereas our approach reflects the limited opportunities to exploit an inflation-unemployment tradeoff in the long run in the St. Louis model.

II. A Rational Expectations Approach

The preceding analysis has some rather serious drawbacks, particularly if expectations are formed rationally. Lucas has argued against traditional policy analysis on the grounds that parameters of "structural" equations in macroeconomic models are unlikely to be invariant to changes in policy rules. This concern is of special importance here, since the hypothesis of differing party regimes is central to the analysis. If parties truly have systematically different policy rules, agents should come to anticipate such differences.

In this section we discuss a preliminary empirical analysis of the effects of such party differences in a model with rational expectations. In the model employed, the effects of alternative money growth rates cannot be viewed in isolation from agents' expectations of money growth, and agents' expectations are assumed to take into account systematic party differences in policy rules. The key elements of the model are a policy rule (embodying party differences) for monetary growth, and an aggregate supply function, that relates deviations of real output from trend values to deviations of actual from expected rates of money growth. In addition, an Okun's law-type equation relates deviations of output from trend to similar deviations in the unemployment rate.

In this model the policy rule specifies that money growth depends on its own lagged value and an intercept which differs for the two parties. Agents are assumed to know the policy rule and to use it in forecasting money growth. However, agents' forecasts are complicated by the occurrence of elections. Election outcomes are not known in advance,

and this leads to forecasting errors when the horizon of a forecast is beyond the date of the next election. When looking beyond the next election, agents are assumed to forecast money growth rates conditional, respectively, upon Democratic and Republican victories, and then to weight the two forecasts by assessed probabilities of victory of the parties. In our empirical analysis, we have computed one- and two-year-ahead election-win probabilities for the parties based on a regression analysis relating observed presidential vote to poll data from preceding years. We then have used these probabilities, along with estimates of the policy rule, to compute one- and two-year-ahead forecasts of money growth rates. These become the expected money growth rates that are used in estimating the aggregate supply function.

Our supply function is similar to that used by Stanley Fischer (1980). A key assumption underlying this supply function is that multiperiod labor contracts lead to the absence of instantaneous labor market clearing, which in turn permits anticipated money growth variations to have real impacts. We thus avoid ruling out real party differences a priori. The supply function specifies that differences between actual money growth and one- and two-year-ahead forecasts of money growth explain deviations of output from trend values. One- and two-year lags of these expectational errors are also included in the aggregate supply function, following Fischer, and Robert Barro and Mark Rush (1980). We have used annual data from 1949 to 1984 to estimate the policy rule, and annual data from 1957 to 1984 to estimate the supply function and the Okun's law equation.

For the analysis of party differences, the key feature of this model is that elections systematically produce money growth rate forecast errors, that in turn produce systematic party differences in real variables. These errors occur even when parties follow their respective rules without deviation. Because expected money growth is a probability weighted average of growth rates anticipated for Democrats and Republicans, actual money growth rates will tend to be higher than expected if Democrats actually win the election and lower than expected if Republicans actually win the election. Our task is to assess the size of these party-related and electorally induced forecast errors, and to estimate the size of the resulting impacts on output and unemployment.

To do so, we simulated with the model under alternative scenarios regarding presidential election outcomes, assuming that parties followed the policy rules we have estimated for them. The simulations indicated that counterfactually altering an election outcome (without changing the a priori probabilities of election for the two parties) typically leads to party differences in the unemployment rate which reach a maximum value of about one-half of a point (with lower unemployment associated with Democratic victory) in the second and third years after an election. Of course, the actual impact of party on deviations of output and unemployment from trends depends on the extent to which an election outcome is a surprise—the bigger the surprise, the bigger the induced deviations from trend. By the fourth year after an election the effect of party is largely dissipated.

The smaller initial unemployment rate difference (compared to the St. Louis model findings) clearly results partly from differences in specification of the model's supply side, but a smaller estimated party difference in the monetary policy rule is also partly responsible. The relative lack of persistence in party impacts is primarily a result of the short lag structure assumed for the money growth forecast errors within the supply function. Future work investigating the robustness of these results under alternative assumptions about the structural features of the model is certainly called for. In any case, while these effects are small in comparison to previous estimates, they are still nontrivial.

The model has some implications regarding political business cycles. In particular, the conventional political business cycle pattern, with early term recession followed by faster growth at the end of a term, is observed under Republicans in this model. A reverse pattern is predicted for Democrats. This may help to explain the casual observa-

tion that the two episodes that best fit the political business cycle model have come under Republicans (Nixon, 1969–72; Reagan, 1981–84).

One institutional change that could dampen fluctuations associated with elections would be to lengthen the lame-duck period between the date of an election and the date of inauguration. In our model, forecast errors from as far back as two years prior to the election can create real disturbances. Thus, if the lame-duck period was lengthened to more than two years, agents could always predict party in power accurately over a two-year horizon. Expectational errors associated with elections could be eliminated, as would the real fluctuations associated with them. Obviously long lame-duck periods may have other costs that could more than offset any gains from economic stabilization.

III. Conclusions

Our findings admittedly rest on limited data, crude and preliminary procedures, and simple models. But the finding of systematic and nontrivial party differences in real economic outcomes has clearly survived an analysis that imposes some constraints from macroeconomic theories, and would seem to warrant further attention.

REFERENCES

Barro, Robert and Rush, Mark, "Unanticipated Money and Economic Activity," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Beck, Nathaniel, "Parties, Administrations, and American Macroeconomic Outcomes," *American Political Science Review*, March 1982, *76*, 83–93.

Fischer, Stanley, "On Activist Monetary Policy with Rational Expectations," in his *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Hibbs, Douglas, "Political Parties and Macroeconomic Policy," *American Political Science Review*, December 1977, *71*, 1467–87.

Policy Responses to Exogenous Shocks in Developing Countries

By BELA BALASSA*

This paper reports on the results of research on the policy responses of developing countries to exogenous (external) shocks in the 1973–78 and 1978–83 periods. These shocks included terms-of-trade effects, associated largely with increases in oil prices; export volume effects, resulting from the recession-induced slowdown of world trade; and, in the second period, interest rate effects, due to increases in interest rates in world financial markets. Policy responses to external shocks took the form of additional net external financing, represented by increased borrowing compared with past trends; export promotion, reflected by increases in export market shares; import substitution, expressed by decreases in the income elasticity of import demand; and deflationary macroeconomic policies, entailing a decline in the growth of demand for imports (For a description of the methodology applied and results for earlier periods, see my 1985 paper).

Table 1 provides summary data on the balance-of-payments effects of external shocks and of policy responses to these shocks in the two periods. Developing countries were classified as outward- and inward-oriented, depending on whether they provided similar incentives to exports and to import substitution or discriminated in favor of import substitution and against exports.

TABLE 1—BALANCE OF PAYMENTS EFFECTS OF EXTERNAL SHOCKS AND OF POLICY RESPONSES TO THESE SHOCKS[a]

	Outward-Oriented Countries		Inward-Oriented Countries	
	1974–78	1979–83	1974–78	1979–83
External Shocks[b]				
Terms of Trade Effects	6.3	8.4	3.6	2.8
Export Volume Effects	2.4	4.9	0.9	0.4
Interest Rate Effects	–	1.7	–	1.6
Total	8.8	15.0	4.5	5.0
Policy Responses[c]				
Additional Net External Financing	−26.4	−11.5	89.0	37.6
Export Promotion	48.6	29.0	−14.9	11.5[d]
Import Substitution	58.5	24.5	15.4	9.8
Effects of Deflationary Policy	19.4	58.0	10.5	41.1

Source: World Bank data base.
[a]For definitions, see text.
[b]Shown as percent of *GNP*.
[c]Shown as percent of External Shocks.
[d] −2.3 excluding fuel exports.

Both groups include newly industrializing countries (*NICs*) and less developed countries (*LDCs*) although, to save space, only their combined results are reported in the table.

Among the *NICs*, Korea, Singapore, and Taiwan adopted an outward-oriented development strategy in the early 1960's and continued with this strategy after 1973. In the mid-1970's, they were joined by Chile and Uruguay, who had previously applied inward-oriented policies but turned outward in response to the external shocks. Conversely, Argentina, Brazil, Israel, Mexico, Portugal, Turkey, and Yugoslavia maintained, or reinforced, their inward-oriented stance. Among

†*Discussants*: Alberto Giovannini, Columbia University, and Nathaniel Leff, Columbia University.

*The Johns Hopkins University and World Bank, 1818 H Street, NW, Washington, D.C. 20433. Research assistance by Shigeru Akiyama is gratefully acknowledged. I alone am responsible for the opinions expressed herein, they do not reflect the views of the World Bank.

the *LDC*s, Kenya, Mauritius, Thailand, and Tunisia were classified as having followed outward-oriented policies, and Egypt, India, Jamaica, Morocco, Peru, Philippines, Tanzania, and Zambia as having pursued inward-oriented policies.

The classification scheme was established for the first period of external shocks; for reasons of comparability, it was retained for the second period even though policy changes were made in several countries. In particular, Turkey undertook a far-reaching policy reform in January 1980, while Chile and Uruguay distorted the system of incentives by failing to adjust their exchange rates pari passu with domestic inflation.

I. External Shocks and Policy Responses to the Shocks in 1973–78 and 1978–83

The results show that outward-oriented countries (*OOC*s) suffered substantially larger terms of trade losses and adverse export volume effects than inward-oriented countries (*IOC*s) during both periods of external shocks. This is explained by the larger share of foreign trade in their gross national product (28 percent of the *OOC*s, on average, in 1973 compared with 10 percent for the *IOC*s) that was only partially compensated by the favorable commodity composition of their exports.

One also observes considerable differences in policy responses to external shocks in the two groups of countries, when the sequencing of these responses is of further interest. In the first period of external shocks, the *IOC*s offset nearly the entire adverse balance-of-payments impact of external shocks by additional net external financing. This was done with a view to maintaining past economic growth rates, notwithstanding the deterioration of the external environment. They did not succeed in this effort, however, and the rate of growth of *GNP* declined during the period.

The lack of output-increasing (expenditure-switching) policies importantly contributed to the deceleration of the rate of economic growth in the *IOC*s. Losses in export market shares practically offset import substitution in these countries, with their combined impact on the balance of payments and on domestic output being virtually nil. At the same time, losses in export market shares accentuated the effects of external borrowing on debt-service ratios, defined as the ratio of net interest payments and amortization to merchandise exports. This ratio nearly doubled in the space of five years, rising from 22 percent in 1973 to 43 percent in 1978, on the average.

The *OOC*s initially applied deflationary policies to limit reliance on external finance so that their debt-service ratio remained at slightly below 12 percent. The resulting decline in *GNP* growth rates remained temporary, however, as the *OOC*s adopted output-increasing policies of export promotion and import substitution that fully compensated for the adverse balance-of-payments effects of external shocks and led to the acceleration of economic growth.

The second period of external shocks thus found the *IOC*s (but not the *OOC*s) with considerable foreign indebtedness. Additional borrowing was possible for a while, except for Turkey that was practically bankrupt in 1979 and Yugoslavia that encountered borrowing limitations in 1980. However, with further increases in their debt-service ratios, the other *IOC*s also approached fiduciary limits and, following the August 1982 Mexican debt crisis, they ceased to be creditworthy for commercial bank loans.

Correspondingly, the *IOC*s made less use of additional net external financing in the second period of external shocks than in the first. They applied deflationary policies instead, leading to a decline in their economic growth rates. This result reflects the fact that the *IOC*s largely eschewed output-increasing policies of export promotion and import substitution.

In fact, the extent of import substitution in the *IOC*s declined during the second period of external shocks and discoveries of oil deposits in Mexico and Peru fully account for the observed increases in average export market shares. Thus, adjusting for the rise in petroleum exports, the *IOC*s again experi-

enced losses in foreign markets. And although the losses were smaller than in the first period of external shocks, this was due to the improved performance of a few countries. The January 1980 policy reform, representing increased outward orientation, led to a near-doubling of Turkish exports between 1980 and 1983, while export subsidies contributed to the expansion of exports in Brazil. All other *IOC*s lost export market shares.

The *OOC*s also applied deflationary policies in response to the external shocks they suffered after 1978. But the resulting decline in *GNP* growth rates again remained temporary and the countries in question subsequently resumed higher rates of economic growth. This occurred as the *OOC*s continued to apply output-increasing policies, leading to increases in export market shares and import substitution, even though they had to rely to a greater extent on deflationary policies than during the first period, when the balance-of-payments effects of external shocks were much smaller.

At the same time, the *OOC*s continued to limit reliance on external finance, so that their average debt service ratio remained below 14 percent notwithstanding increases in world interest rates. By contrast, debt-service ratios continued to rise in the *IOC*s, albeit at a slower rate than beforehand, reaching 53 percent in 1983.

II. The Policy Measures Applied and their Economic Effects

In both periods, then, the *OOC*s made considerable gains in export market shares. In turn, the *IOC*s experienced losses in market shares, even though these losses were attenuated in the second period of external shocks by the export-promoting measures applied in a few countries. Providing similar incentives to exports and to import substitution in the *OOC*s, compared with the continued bias of the incentive system against exports in the *IOC*s, importantly contributed to the observed differences in export performance.

Another contributing factor was exchange rate policy, with the adoption of realistic exchange rates in the *OOC*s and appreciation in real (inflation-adjusted) terms in most of the *IOC*s. Increased overvaluation in the *IOC*s was associated with foreign borrowing that obstructed adjustment in the exchange rate as the external financing of the balance-of-payments deficit permitted maintaining an overvalued currency. In turn, the *OOC*s did not use foreign borrowing to support the exchange rate.

The *OOC*s also experienced import substitution to a greater extent than the *IOC*s. This result may appear surprising since the bias against exports favored the replacement of imports by domestic production in the latter group of countries. Various factors contributed to this outcome.

To begin with, the adoption of realistic exchange rates contributed to import substitution parallel with export expansion in the *OOC*s, which was not the case in the *IOC*s. Export expansion in the *OOC*s also permitted simultaneous import substitution as the exploitation of economies of scale led to lower costs. Such efficient import substitution contrasted with inefficient import replacement in many of the *IOC*s, where net foreign exchange savings tended to decline as shifts occurred towards industries where the countries in question had a comparative disadvantage and increasingly encountered domestic market limitations.

Furthermore, the *OOC*s experienced import substitution in the primary sector as they provided similar incentives to primary and to manufacturing activities, while primary production suffered considerable discrimination in the *IOC*s. Finally, the former, but not the latter, group of countries encouraged energy savings, representing import substitution in fuels under the conventions adopted in this study, by increasing energy prices parallel with the rise in world market prices.

The lack of discrimination against exports and against primary activities raised the level of investment efficiency in the *IOC*s, thereby contributing to their economic growth. The liberalization of prices and the application of economic considerations in public investment projects also had a favorable impact on

the efficiency of investment in these countries. Export expansion, too, had beneficial effects by permitting higher capacity utilization and the exploitation of economies of scale.

In turn, the bias of the incentive system against exports and against primary activities, together with the widespread application of price control, reduced the efficiency of investment in the *IOC*s. The situation was aggravated by the lack of sufficient attention given to economic considerations in the large public investment programs of these countries.

These considerations explain the observed differences in incremental capital-output ratios, taken to represent the level of investment efficiency notwithstanding the well-known limitations of these ratios. In the 1973–79 period, the ratios averaged 4.1 in the *OOC*s and 4.9 in the *IOC*s; the corresponding figures were 7.4 and 8.6 in the 1979–84 period, when the deflationary policies applied raised the ratios in both groups of countries.

The *OOC*s also exhibited higher domestic savings ratios than the *IOC*s. Between 1973 and 1979, these ratios averaged 25.6 percent in the former group of countries and 21.0 percent in the latter. The differences were maintained in the 1979–84 period, the average ratios being 25.7 percent in the *OOC*s and 20.9 percent in the *IOC*s.

These differences pertain equally to public and to private savings. While the *IOC*s practiced public dissaving as they incurred large budget deficits, the *OOC*s limited the size of these deficits. Also, real interest rates tended to be negative in the *IOC*s and positive in the *OOC*s, with corresponding effects on private savings.

Higher investment efficiency and higher domestic savings ratios in the *OOC*s were only partially offset by greater foreign borrowing in the *IOC*s. Correspondingly, rates of economic growth were considerably higher in the former than in the latter group of countries, with the differences increasing over time.

III. Conclusions

This paper reviewed the adjustment experience of developing countries applying different policies in response to the external shocks of the 1973–78 and 1978–83 periods. Although outward-oriented countries suffered considerably larger external shocks than inward-oriented countries, these differences were offset several times as a result of the policies followed. Thus, while the *OOC*s accepted a temporary decline in *GNP* growth rates in both periods in order to limit reliance to foreign borrowing, their economic growth accelerated subsequently, owing to the output-increasing policies applied.

The *IOC*s relied practically exclusively on foreign borrowing in response to the external shocks of the first period. But, the bias of the incentive system against exports and primary activities, price control, and the frequent choice of high-cost public investment projects did not provide for the efficient use of these funds, and of investable funds in general, leading to lower economic growth rates and compromising their creditworthiness.

In eschewing output-increasing policies, limitations on external finance in the second period of external shocks led to the application of deflationary policies in the *IOC*s, further increasing differences in the growth performances of the two groups of countries. Between 1982 and 1984, *GNP* growth rates averaged 5.3 percent in the *OOC*s and 1.7 percent in the *IOC*s.

REFERENCE

Balassa, Bela, "Adjustment Policies in Developing Countries: A Reassessment," *World Development*, reprinted as Essay 5 in my *Change and Challenge in the World Economy*, London: Macmillan, 1985, 89–101.

Monetary Policy Responses to Exogenous Shocks

By MAXWELL J. FRY AND DAVID M. LILIEN*

The oil-price shocks of 1973–74 and 1979–80 reduced output growth in oil-importing countries. Using monetary policy to accommodate exogenous shocks of this kind undoubtedly works. But the more such monetary policy is used, the less effective it becomes. And discretionary monetary policy has a negative effect on economic growth in the long run. This is how we interpret the econometric results reported below.

We find that money is neutral neither in the medium term nor in the long run. The effects of current and lagged money growth shocks on output growth are significantly positive. Over time, however, discretionary monetary policy creates a higher variance of money growth shocks. The period-average variance of money growth shocks together with our indicator of an accommodative monetary policy regime are both negatively related to the rate of growth in real gross domestic product (*GDP*).

Higher variance of money growth shocks reduces both the medium-term impact of discretionary monetary policy on output growth and output growth itself in the long run (see Robert Lucas, 1973, and Roger Kormendi and Philip Meguire, 1984, 1985). The only way of testing both the medium- and long-run effects of discretionary monetary policy is by pooling time-series data across countries; we use 647 observations for 55 developed and developing countries. We believe this to be the first appropriate test.

Monetary accommodation of exogenous shocks, specifically accommodation of the 1973–74 oil-price increase, works temporarily to offset the output growth-reducing impact of the shock. However, accommodation adds noise to the economic environment and reduces output growth in the longer run. Indeed, we find that our monetary accommodation variable performs in virtually the same way as the variance of money growth shocks.

Expansionary fiscal policy has medium- and long-run effects on output growth that are similar to monetary accommodation. Specifically, expansionary fiscal policy raises the ratio of net government credit to total domestic credit in countries lacking well-developed direct financial markets. In the medium term, a positive government credit shock raises output growth by stimulating aggregate demand. But a higher government credit ratio lowers output growth in the long run by starving the private sector of finance for productive investment.

I. The Model

Country i's rate of growth in *GDP* in year t is composed of a country-specific long-run or normal growth rate G_i^n and a country-specific cyclical growth rate G_{it}^c:

$$G_{it} = G_i^n + G_{it}^c. \tag{1}$$

Long-run output growth is a function of country-characteristic dummy variables and a number of variables designed to capture the long-run effects of monetary policy:

$$\begin{aligned} G_i^n = \alpha_0 &+ \alpha_1 IND_i + \alpha_2 OILX_i \\ &+ \alpha_3 DM_i^* + \alpha_4 MVAR_i \\ &+ \alpha_5 MACC74_i + \alpha_6 DP_i^* + \alpha_7 NDCG_i^*, \end{aligned} \tag{2}$$

where IND_i and $OILX_i$ are dummy variables for industrialized and oil-exporting countries. Industrialized countries may experience lower output growth rates ($\alpha_1 < 0$) than developing countries until per capita incomes in *LDC*s catch up (see Robert Barro, 1984, ch. 12). The variable DM_i^* is the long-run rate of growth in the money supply, mea-

*School of Social Sciences, University of California, Irvine, CA 92717. We thank Carol Jackson for installing the *IFS* tape at UCI, Jeri Fender for extracting from it the data used in this study, and Kevin Lang, Nathaniel Leff, Edward Shaw, and Kenneth Small for comments.

sured as a simple average of the annual $M1$ growth rates over the sample period. Money neutrality requires α_3 to be zero. We test rather than impose this restriction.

Two other measures of monetary policy are included in the equation. $MVAR_i$ is the variance of money growth shocks (described below) and is a measure of the variability and uncertainty of monetary policy. Constantine Glezakos (1978), Axel Leijonhufvud (1981), and others show that greater uncertainty with respect to the future price level reduces output growth.

The variable $MACC74_i$ is an indicator of monetary accommodation of the 1973–74 oil shock. It takes a value of one if the ratio of $M1$ to nominal GDP rose between 1974 and 1977. While one would not expect accommodation of a single oil shock to affect the long-run rate of economic growth, we take $MACC74_i$ to indicate a country's willingness to use monetary policy to accommodate other supply shocks. Few would doubt that monetary accommodation can reduce the negative impact of supply shocks on output growth in the short run. Here, however, we are measuring its impact on long-run growth. Since $MACC74_i$ could be an alternative measure of variability in monetary policy, we expect it to lower the rate of economic growth in the long run.

The period-average inflation rate is denoted DP_i^*. Inflation reduces long-run output growth in the inside-money model of Fry (1980, 1982), but increases growth in James Tobin's (1965) outside-money model. Since DP_i^* may be determined simultaneously with G_i^n, we also estimate a reduced-form model in which DP_i^* is omitted.

The variable $NDCG_i^*$ is the trend ratio of net government credit to total domestic credit. When government extracts greater seigniorage by increasing the proportion of domestic credit allocated to the public sector, credit availability for the private sector is reduced. Alan Blinder and Joseph Stiglitz (1983), and Fry (1980, 1981) analyze how this credit availability effect retards output growth.

Country i's cyclical output growth in period t is a function of both policy shocks and exogenous supply shocks:

$$(3)\quad G_{it}^c = \beta_{i1}DMR_{it} + \beta_{i2}DMR_{it-1} + \beta_{i3}DMR_{it-2} + \lambda_{i1}OILDP_t + \lambda_{i2}OILDP_{t-1} + \lambda_{i3}OILDP_{t-2} + \theta_1 NDCGR_{it} + \gamma_1 G_{it-1}^c + \gamma_2 G_{it-2}^c,$$

where DMR_{it-j} represents innovations to the time-series process of money growth. It is the residual of country-specific regressions of the money growth rate on its own lagged value and a time trend. This variable is similar to that used by Barro (1978) and others to measure unanticipated money. Since we are measuring output as a growth rate, long-run neutrality requires that the sum of the β coefficients equal zero. Short-run neutrality or total policy ineffectiveness implies that each β would equal zero.

Note that we do not assume money growth shocks have the same effect in all countries; the βs have i subscripts. Rather we test the Lucas hypothesis that money growth shocks have a smaller impact on real variables in countries where highly variable monetary policy makes the current money growth rate a poor signal of real disturbances:

$$(4)\quad \beta_{ij} = \delta_{0j} + \delta_{1j}MVAR_i + \delta_{2j}MACC74_i.$$

Following Lucas, we expect $\delta_1 < 0$ and $\delta_2 < 0$.

Supply shocks are measured by the rate of change in oil prices $OILDP_t$. We allow oil shocks to have an impact on oil-exporting countries that differs from their impact on oil-importing countries:

$$(5)\qquad \lambda_{ij} = \varphi_{0j} + \varphi_{1j}OILX_i.$$

We also tested other measures of supply shocks, including country-specific changes in the terms of trade. The coefficients had the expected signs but were less significant than the oil-price coefficients.

Finally, we include unanticipated changes in the government credit ratio $NDCGR_{it}$

TABLE 1—OUTPUT GROWTH RATE EQUATIONS[a]

Variable	Long Run		
Constant	.037	.027	.026
	(.004)	(.004)	(.004)
DM_i^*	.120	.125	.043
	(.023)	(.023)	(.021)
$MVAR_i$	−.213	−.316	−.330
	(.147)	(.145)	(.150)
$MACC74_i$	−.007		
	(.002)		
DP_i^*	−.081	−.077	
	(.011)	(.011)	
$NDCG_i^*$	−.011		
	(.004)		
IND_i	−.007	−.004	−.004
	(.003)	(.003)	(.003)
$OILX_i$	.001	.002	.004
	(.005)	(.004)	(.005)
	Medium Term		
DMR_{it}	.168	.161	.145
	(.023)	(.023)	(.024)
DMR_{it-1}	.104	.097	.067
	(.024)	(.024)	(.025)
DMR_{it-2}	−.035	−.044	−.056
	(.024)	(.024)	(.025)
$MVAR_i \cdot DMR_{it}$	−.012	−.012	−.020
	(.004)	(.004)	(.004)
$MVAR_i \cdot DMR_{it-1}$	−.004	−.003	−.007
	(.004)	(.004)	(.004)
$MVAR_i \cdot DMR_{it-2}$	.006	.006	.005
	(.003)	(.003)	(.004)
$NDCGR_{it}$	.013		
	(.006)		
$OILDP_t$	−.020	−.020	−.027
	(.006)	(.006)	(.006)
$OILDP_{t-1}$	−.015	−.014	−.017
	(.006)	(.006)	(.007)
$OILDP_{t-2}$	−.003	−.001	−.004
	(.006)	(.006)	(.006)
$OILX_i \cdot OILDP_t$	.027	.025	.026
	(.017)	(.016)	(.017)
$OILX_i \cdot OILDP_{t-1}$	−.017	−.018	−.012
	(.017)	(.017)	(.017)
G_{it-1}	.179	.210	.287
	(.043)	(.042)	(.043)
G_{it-2}	.088	.113	.140
	(.042)	(.041)	(.042)
R^2	.314	.298	.245

[a]Standard errors are shown in parentheses.

(measured in the same way as DMR_{it}) and lagged output growth rates in equation (3).

The regressions reported in Table 1 are estimates of the equation for real GDP growth G_{it} derived from equations (1) to (5).

II. The Estimates

Annual data for 1950–83 came from the *International Financial Statistics* October 1985 computer tape. The country sample consists of all members of the International Monetary Fund with populations over two million and complete data sets for 1960–83. In order to reduce noise caused by substantial year-to-year fluctuations in agricultural output, we took two-year averages of all the raw annual data. Hence variables that are included with zero, one- and two-period lags use data spanning six years. Two-year averaging means that we are not analyzing short-run behavior, for which monthly or quarterly data would be needed, but rather medium- and long-run behavior.

Three pooled time-series GDP growth estimates are shown in Table 1. Long-run output growth is positively and significantly associated with the average rate of money growth. We reject (long-run) neutrality with respect to DM_i^* at the 99 percent level. We also confirm the hypothesis that money variability as measured by $MVAR_i$ and $MACC74_i$ reduces long-run output growth.

Care is needed in interpreting these results. While higher money growth significantly increases output growth, most countries with high values of DM_i^* also have high values of $MVAR_i$ that offset this effect. Furthermore, higher inflation produced by higher money growth itself reduces long-run output growth.

There is clearly an important question about the interpretation of $MACC74_i$. A change in the ratio of money to nominal GDP between 1974 and 1977 could well be a result of a change in the opportunity cost of holding money rather than a change in monetary policy stance. We interpret it to be a short-run monetary policy indicator, as do Bela Balassa and Desmond McCarthy (1984), for two reasons. First, the change in the money-GDP ratio over this period is strongly correlated with the change in the ratio of net government credit to GDP. Second, we obtain similar results from several alternative measures of monetary accommodation, including the change in the ratio of net govern-

ment credit to *GDP* and the sign of the coefficient of terms-of-trade growth in a money growth equation.

The estimates show that inflation and the ratio of net government credit to total domestic credit reduce output growth in the long run. With zero-interest-earning required reserves, higher inflation reduces the real return on all forms of money balances. The discriminatory tax on financial intermediation imposed by the reserve requirement increases as inflation and nominal interest rates rise. By reducing the attractiveness of the financial sector's liabilities, the reserve-requirement tax reduces the relative size of this sector of the economy. Hence, the private sector suffers a credit squeeze in real terms as inflation (and nominal credit expansion) accelerates. When government increases the proportion of domestic credit allocated to the public sector, credit availability for the private sector is again reduced. Suboptimal provision of institutional credit lowers both the quality and quantity of investment, and hence also the rate of growth in output.

Medium-term output growth is positively and significantly affected by money growth and net government credit shocks, as well as lagged output growth rates. An expanded table of results (available upon request) shows that dropping the lagged dependent variable increases the other coefficient values and their statistical significance somewhat but reduces the overall fit of the equation.

Output growth is negatively and significantly affected in the medium term by the rate of change in oil prices. Note that the immediate effect of an oil shock in oil-exporting countries is a modest increase in output growth (measured by the sum of the coefficients of $OILDP_t$ and $OILX_i \cdot OILDP_t$). After two years (one period), however, growth is depressed even in oil-exporting countries, as predicted by "Dutch disease" models.

The positive effect of money growth shocks on output growth has been widely documented by Barro (1978), Kormendi and Meguire (1984), and others. Since the sum of the DMR_{it-j} coefficients is significantly positive, we reject neutrality (sum of coefficients equal to zero) at the 99 percent level in all of equations that include the current and two lagged values of DMR_{it-j}, and at the 95 percent level in equations (not shown) containing three lagged values of DMR_{it-j}. While we ourselves are not completely comfortable with this result, the estimates imply that money growth shocks have a permanent effect on the *level* of output.

Lucas and Kormendi and Meguire (1984) show that the coefficients of money growth shocks in country-specific output growth equations are smaller the greater the variance of the shocks. Here we find the same phenomenon in a single pooled time-series model. The impact of money growth shocks on output growth is reduced by greater variance of the shocks, as shown by the negative coefficients of $MVAR_i \cdot DMR_{it-j}$. This effect is large enough to turn the implied total coefficient of DMR_{it-j} negative (although not significantly) for several of the high $MVAR_i$ countries in our sample. This means that increasing the money supply can raise output growth in the medium term only when such a policy is used infrequently. A history of accommodation and highly variable monetary policy makes monetary policy totally ineffective in the medium term and, as we have already noted, leads to reduced output growth in the long run.

Shocks to the ratio of net government credit to total domestic credit increase output growth over and above the indirect effects of such shocks produced by concomitant money growth shocks. Government deficits in most developing countries are financed predominantly by the central and commercial banks. Hence an increased deficit shows up as a shock to the government credit ratio. Since government credit is expanded by increasing money, we interpret the positive coefficient of the government credit shocks to be pure fiscal stimulus; higher government spending raises output growth temporarily.

REFERENCES

Balassa, Bela and McCarthy, F. Desmond, "Adjustment Policies in Developing Countries, 1979–83," Staff Working Paper No. 675, World Bank, November 1984.

Barro, Robert J., "Unanticipated Money, Out-

put, and the Price Level in the United States," *Journal of Political Economy*, August 1978, *86*, 549–80.

______, *Macroeconomics*, New York: Wiley & Sons, 1984.

Blinder, Alan S. and Stiglitz, Joseph E., "Money, Credit Constraints, and Economic Activity," *American Economic Review Proceedings*, May 1983, *73*, 297–302.

Fry, Maxwell J., "Money, Interest, Inflation and Growth in Turkey," *Journal of Monetary Economics*, October 1980, *6*, 535–45.

______, "Inflation and Economic Growth in Pacific Basin Developing Economies," *Federal Reserve Bank of San Francisco Economic Review*, Fall 1981, 8–18.

______, "Analysing Disequilibrium Interest-Rate Systems in Developing Countries," *World Development*, December 1982, *10*, 1049–57.

Glezakos, Constantine, "Inflation and Growth: A Reconsideration of the Evidence from *LDCs*," *Journal of Developing Areas*, January 1978, *12*, 171–82.

Kormendi, Roger C. and Meguire, Philip G., "Cross-Regime Evidence of Macroeconomic Rationality," *Journal of Political Economy*, October 1984, *92*, 875–908.

______ **and** ______, "Macroeconomic Determinants of Growth: Cross-Country Evidence," *Journal of Monetary Economics*, September 1985, *16*, 141–63.

Leijonhufvud, Axel, "Costs and Consequences of Inflation," in his *Information and Coordination: Essays in Macroeconomic Theory*, New York: Oxford University Press, 1981, 227–69.

Lucas, Robert E., Jr., "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

Tobin, James, "Money and Economic Growth," *Econometrica*, October 1965, *33*, 671–84.

Developing Country Exchange Rate Policy Responses to Exogenous Shocks

By MOHSIN S. KHAN*

The late 1970's and early 1980's proved to be extremely trying economic times for the developing countries. Throughout most of the period, a combination of exogenous shocks, such as worsening terms of trade, falling growth rates in industrial countries, and sharp changes in the cost and availability of foreign financing, created serious macroeconomic management problems for policymakers in these countries. Adjustment to these shocks required fiscal and monetary restraint to control both public and private spending, and the adoption of a flexible exchange rate policy to prevent the emergence of unsustainable current account deficits, growing foreign debt burdens, and steady losses of international competitiveness. With certain exceptions, developing countries generally did not follow this policy prescription, and consequently compounded the negative effects of the exogenous shocks.

The purpose of this paper is to evaluate the exchange rate responses of developing countries to the variety of exogenous shocks they faced in recent years. Essentially this involves an analysis of the behavior of the real exchange rate. Exchange rate policy responses have to be judged in terms of how the authorities used combinations of nominal exchange rate action and other policies to restrain domestic prices and factors to either support or offset the movements in the real exchange rate caused by external shocks. In this paper I discuss first the principal external shocks that occurred during the past decade, and then the likely effects of these on the real exchange rate. The picture is completed by a description of how real exchange rates actually evolved in developing countries during the period under consideration.

*IMF and the World Bank, 1818 H Street, NW, Washington, D.C. 20433. I am grateful to Tom Walter for excellent research assistance. The views expressed are my sole responsibility.

I. External Shocks and Real Exchange Rates

A. *Pattern of External Shocks*

In a recent paper, Malcolm Knight and I (1983) identified these external factors as being mainly responsible for the recent current account difficulties experienced by non-oil developing countries during the 1970's: the deterioration in the terms of trade, the slowdown in economic activity in the industrial world, and, towards the end of the period, the sharp rise in real interest rates in international capital markets. To this list we can now add the drastic contraction in external financing in 1982–84.

Considering the period 1977–84, the data reported in IMF (1985a) show that the terms of trade of non-oil developing countries deteriorated at an annual average rate of 1.2 percent. There was a modest improvement of almost 3 percent in 1983–84, but this was hardly sufficient to compensate for the cumulative losses that occurred since 1977. Insofar as growth in industrial countries is concerned, during 1977–79 real *GNP* of the industrial countries grew at an annual average rate of about 4 percent, which helped to moderate the effects of worsening terms of trade on the developing countries. The average growth rate fell dramatically, however, in the period 1980–82 to less than 1 percent per year. Growth picked up once again in 1983–84 and averaged nearly 4 percent per year. After having remained strongly negative for a number of years, foreign real interest rates, defined as the nominal foreign interest rate adjusted for the percentage change in export prices of oil-importing developing countries, started to rise in 1978. The foreign real interest rate climbed to an average of

9 percent a year during 1978–84, and between 1981 and 1984 reached an annual average of about 16 percent. This increase forced a number of developing countries into serious debt-servicing difficulties, particularly towards the end of the period when recourse to foreign financing diminished.

Between 1977 and 1981 net external borrowings by oil-importing countries nearly tripled from \$28 billion to over \$84 billion. Latin American countries were the primary recipients of these flows, increasing their borrowing from about \$17 billion in 1977 to around \$56 billion in 1981. In 1982, the year of the Mexican crisis, the ability of developing countries to borrow abroad began to weaken. By 1983 net foreign financing for Latin America fell to near zero, and for all oil-importing developing countries amounted to only \$19 billion. This situation was repeated for 1984 where net foreign borrowings by these countries declined by a further \$5–6 billion.

B. *Theoretical Effects of External Shocks on the Real Exchange Rate*

The likely effects on the real exchange rates of developing countries of the types of shocks described above can be analyzed within the framework developed by Rudiger Dornbusch (1985) to study the welfare costs associated with external shocks. The economy is assumed to produce nontraded goods and two types of exportables—manufactures and primary commodities. It also imports manufactures. The country is small so that the world price of primary commodities and the price of imported manufactures are treated as given. The terms of trade is defined as the ratio of the price of commodities to the price of imported manufactures, and the real exchange rate as the relative price of nontraded goods to imported manufactures.

Suppose there is a decline in the relative price of commodities, which in the context of the model is the same as a worsening of the terms of trade. This deterioration in the terms of trade has both supply- and demand-side effects, since it lowers factor costs and increases production of domestic manufactures and nontraded goods, and at the same time reduces real income and thus spending on these goods. Consequently, in this flexible-price model the relative price of nontraded goods has to fall and there is a depreciation of the real exchange rate. A decline in foreign real income would create an excess supply of manufactures, and again the relative prices of both manufactures and nontraded goods must fall to maintain equilibrium. An increase in the foreign real interest rate would tend to reduce borrowing and spending through two separate channels. First, it would reduce investment and increase savings. Second, the higher interest rates would raise debt-servicing requirements and this would impact on domestic expenditures as real disposable income would be reduced. Both factors would reduce demand for manufactures and nontraded goods, causing their relative prices to fall, and implying a real depreciation. Conversely, an inflow of capital from abroad would allow domestic residents to raise spending of all goods and there would be a real appreciation. An outflow of capital would simply reverse this process and give rise to a real depreciation.

C. *Real Exchange Rate Behavior in Developing Countries*

We next ask: how did real exchange rates of developing countries evolve over the period that external shocks were evident, and were these movements consistent with the predictions of the theory? The behavior of real effective exchange rates, calculated as an index of a country's prices relative to its trading partners adjusted for exchange rate changes, for different country groups are shown in Figure 1. The method of calculation and the classification of the country groups are described in IMF (1985b).

The first panel of the figure shows the real effective exchange rate indices for both the developing countries (including oil-exporting countries) and the non-oil developing countries groups. Until about 1976 the series moved somewhat differently, with the developing countries group experiencing a small real appreciation. When oil-exporting countries are excluded from the group, the real exchange rate index stays relatively flat for

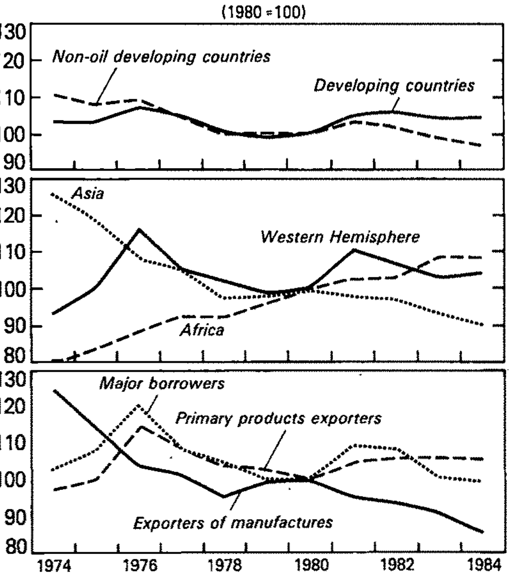


FIGURE 1. DEVELOPING COUNTRIES REAL EFFECTIVE EXCHANGE RATES, 1974–84

the earlier period. From 1977 through 1980, the real exchange rate indices depreciated steadily as the terms of trade of non-oil developing countries worsened. The period 1980–82 was characterized by a continued deterioration in the terms of trade, a sharp slowdown in the growth rates of industrial countries, and the dramatic increase in foreign real interest rates, all of which, other things equal, would have tended to depreciate the real exchange rate. In point of fact, the real exchange rates for developing countries appreciated. This phenomenon can be attributed either to the larger flow of foreign financing or, perhaps more importantly, to inappropriate exchange rate policies. While a number of non-oil developing countries did alter their policies and allowed the real exchange rate to depreciate during 1983–84, the oil-exporting countries experienced a continued appreciation, thus keeping the average index for all developing countries from declining.

Looking more closely at the experiences of the subgroups of countries that make up the overall non-oil developing group, we can see much more diversity in real exchange rate movements. In the second panel of Figure 1, we show the paths of the real exchange rates for countries grouped on a regional basis. The index for the Asian group indicates a continuous real depreciation throughout the period, while the African countries experienced a steady real appreciation, even though they were the most susceptible to variations in the world commodity prices which impacted directly on their terms of trade. The picture for the Western Hemisphere group is more mixed. In the wake of the sharp increase in the prices of oil-related products in 1973–74, the real exchange rate appreciated until around 1976. This was followed by a period of depreciation that was partly due to external developments and partly to active policy. The large-scale foreign borrowing by Latin American countries that financed domestic imbalances led to a sharp real appreciation, amounting to over 10 percent, in 1981. From then on, as strong adjustment efforts were undertaken, including greater flexibility in the management of the exchange rate, the real exchange rate depreciated significantly.

The division of non-oil developing countries by type of export (the third panel in Figure 1) provides additional information on the real exchange responses of developing countries to external shocks. The path for the exporters of manufactures closely mirrors the real exchange rate behavior of Asian countries. Apart from an appreciation in 1979–80, which was a consequence basically of the increase in oil prices, the exporters of manufactures have followed policies to increase their market shares in both the industrial countries and other developing countries. The real exchange rate movements in the case of the primary product exporters are more erratic. Even when primary product prices fell in 1975 the real exchange rate was allowed to appreciate. This process was reversed in the following four years. On average the real exchange rate appreciated by about 5 percent in 1981, and has been relatively stable since then. The real exchange rates of the major borrowing countries are obviously very closely related to the movements in foreign real interest rates. When real interest rates in the international credit markets declined, the

real exchange rate depreciated from 1976 to 1979. The real appreciation from 1980 onwards was also associated with the very large increase in foreign interest rates. Basically this pattern suggests that the authorities followed policies that worked against the effects of external shocks. The adjustment programs adopted when foreign financing fell off in 1983–84 showed up in a depreciation of the real effective exchange rates, amounting to about 10 percent.

II. Conclusions

In summary, while it is difficult to isolate exchange rate policy responses from other policies that were implemented simultaneously, in examining the historical evidence over the last decade or so, one gets the distinct impression that certain developing countries, such as those in southeast Asia, were more successful in their adjustment to external shocks. Others, particularly African primary product exporters, and the major debtor countries, were less inclined to adopt appropriate policies and thus experienced substantial real appreciations of the exchange rate. There were several possible reasons advanced for the lack of adjustment. In some countries the authorities were reluctant to take measures to realize an effective exchange rate depreciation large enough to counteract the effects of external shocks. In others, the authorities may not have felt a strong need for adjustment, owing to such factors as a reasonably strong reserve position, the availability of foreign capital, and perhaps the expectation of an early reversal of the adverse international developments. In the end, however, adjustment was forced on the countries as foreign financing virtually disappeared. Between 1981 and 1984, the trend in developing countries has been away from fixed pegs towards more flexible arrangements, and during this period, of the 106 developing countries for which the requisite data is available, nearly one-half had adopted policies that resulted in depreciating effective exchange rates.

REFERENCES

Dornbusch, Rudiger, "Policy Performance Links between Debtor LDCs and Industrial Nations," *Brookings Papers on Economic Activity*, 2:1985, 303–56.

Khan, Mohsin S. and Knight, Malcolm D., "Determinants of Current Account Balances of Non-Oil Developing Countries in the 1970s: An Empirical Analysis," *IMF Staff Papers*, December 1983, *30*, 819–42.

International Monetary Fund, (1985a) *World Economic Outlook*, Washington: IMF, April 1985.

______, (1985b) *Annual Report 1985*, Washington: IMF, September 1985.

Fiscal Policy Responses to Exogenous Shocks in Developing Countries

By VITO TANZI*

During the past decade, the developing countries have been subjected to various exogenous shocks that have made the pursuit of sound economic policy, and particularly that of sound fiscal policy, very difficult. In this paper I discuss the factors associated with these exogenous shocks; the impact of these shocks on fiscal variables; and some of the policy responses by countries. "Exogenous shocks" are defined as uncontrollable external events that have substantial effects on a country's income level.

I. Factors Associated with External Shocks

The most important exogenous shocks have been the following:

Changes in export earnings: Many developing countries rely heavily on the export of one or few commodities (oil, coffee, copper, etc.) for their foreign exchange earnings. Shocks may originate from unexpected changes in the prices arising from changes in supply conditions or in the level of demand for these commodities. A frost in Brazil, that raises the international price of coffee, raises the foreign exchange earnings of other coffee exporters. An oil embargo by the major Middle Eastern oil-exporting countries had the same effect on the earnings of other oil-exporting countries. Major world booms and recessions, by affecting the level of commodity demand, have generated positive or negative shocks for *LDC* exports.

Changes in major import prices: The most obvious example is provided by oil prices since 1973. In view of the great importance of oil in the imports of many countries, when oil prices rose sharply in the 1970's, the real incomes of many oil-importing countries were significantly reduced.

Changes in the cost of foreign borrowing: As many developing countries are heavy borrowers, a change in the interest rate in international capital markets can be an important exogenous shock. The cost of international borrowing to a given country could also go up because of a changed perception of risk associated with lending to that country. Although the effects on borrowing costs may be the same, the latter is not a truly "exogenous" shock. When the cost of borrowing rises, it affects the cost of new funds as well as the cost of servicing the existing stock of foreign debt. If the size of the debt is high and its maturity is short, the rise in interest expenditure can be substantial. If the foreign debt is mostly public, budgetary expenditures are directly affected.

Changes in the availability of foreign credit: This type of shock is not the same as the previous one. Around 1982, the world witnessed a dramatic reduction in the willingness of commercial banks to lend to many developing countries. Mexico, for example, saw its foreign borrowing fall from $18 billion in 1981 to $5 billion in 1983. The debt crisis made new loans unavailable to many countries, thus reducing their ability to continue financing through this source their current expenditure levels.

Changes in the level of foreign grants: In many countries, and especially in the smaller ones, an important exogenous shock may come in the form of sudden changes in the availability of foreign grants or of concessionary loans. Countries that have relied on these sources for their domestic expenditure will be forced to reduce their spending when those grants are no longer available. Examples of these shocks abound, especially in Africa.

Changes in other factors: Shocks may at times also be associated with factors such as

*Director, Fiscal Affairs Department, International Monetary Fund, Washington, D.C. 20431. The assistance received from Ke-Young Chu and Bassirou A. Sarr is greatly appreciated.

changes in foreign workers' remittances, changes in direct foreign investment, changes in the level of capital outflow by nationals, and so on. In many cases these changes can be traced to the countries' own policies; therefore, they are not genuinely exogenous.

II. Effects of Exogenous Shocks on Fiscal Variables

The factors mentioned above affect not just the incomes of countries but also their fiscal variables. They may improve or worsen the fiscal situation and, by doing so, they may bring about policy responses. The automatic impact of external shocks on the fiscal variables is likely to be much more important in developing countries than in industrial countries. At the same time the ability of developing countries to neutralize these effects, if they wished to do so, is much more limited.

In industrial countries the external shocks affect incomes and economic activity much more than the fiscal variables themselves as the fiscal sector is not closely linked to the external sector. Therefore, the observed changes in the fiscal variables can be attributed to policy responses. For example, when the increase in the oil price in 1974 reduced the real incomes of industrial countries, the governments responded by increasing the level of public spending through transfers to families. This increase in public spending was not automatic but reflected a conscious, discretionary governmental reaction. Apart from the cyclical impact that affected tax revenues, the increase in fiscal deficits in OECD countries in 1975 were policy induced.

In the developing countries, the impact of external shocks on the fiscal variables is much more direct or automatic. Therefore, the observed change in the fiscal variables should not be attributed mainly to policy changes. For this reason, it is very difficult, when dealing with these countries, to isolate the changes in fiscal variables that reflect genuine policy responses from those that reflect automatic effects. Thus, studies that attempt to estimate from observed fiscal changes the fiscal policy response to exogenous shocks are likely to reach misleading conclusions.

The reason for the above conclusion is the close link that exists in *LDC*s between the budget and the foreign sector. This link depends on: (a) the high proportion of foreign trade taxes in total revenue; (b) the high proportion of domestic sales taxes collected from imports; (c) the heavy reliance of corporate income taxes on exports of mineral products; (d) the reliance on the part of the public sector on foreign borrowing or foreign grants; (e) the high proportion of foreign debt that is public; (f) the widespread attempts in these countries to insulate some domestic prices from movements in world prices; and so on.

Foreign trade taxes (import plus export duties) account for more than one-third of the total tax revenue of *LDC*s. This figure, however, does not convey the full importance of the external sector in public revenue as corporate income taxes, which are mostly collected from mineral exports, account for another 18 percent and "domestic" taxes on goods and services are often levied largely on imported goods. More than 50 percent of the tax revenue of developing countries may be directly related to the foreign sector. Furthermore, in many of these countries some of the important export sectors (petroleum, phosphates, bauxite, etc.) are government owned (see my 1986a paper). When the price of those commodities changes, the effect on public revenue can be direct and immediate. Much of the foreign borrowing of *LDC*s is made by the public sector. When the availability or the cost of foreign loans changes, government resources are again immediately and directly affected.

To some extent the same close link between the fiscal sector and the foreign sector exists on the expenditure side. Some government expenditures are financed by earmarked taxes. When tax revenue falls, because of external shocks, the resources available for these expenditures also decline. The size of many subsidies depends on the difference between the international prices of some imported products and their domestic prices. When the international prices increase or the exchange rate appreciates, the amount of the subsidy and thus the budget deficit also increase. Some external shocks have an immediate impact on the financing of invest-

ment expenditure as concessionary loans or grants are often tied to specific projects, so that when these loans or grants change, the resources available for these investments also change.

In conclusion, while shocks affect the levels of real incomes in both industrial and developing countries, they have far more pronounced and direct effects on the fiscal sectors of the *LDC*s.

III. Policy Response

There is some literature that is relevant for assessing what the "optimal" fiscal reaction of developing countries to exogenous shocks should be (see Guido Tabellini, 1985). However, much of this literature is highly theoretical and it assumes that over the short run, policymakers can control the policy instruments; it also assumes that they have the interest and the knowledge to pursue optimal policies. Unfortunately the real world is much more complex. Some obstacles that exist in *all* countries are far more important in *LDC*s.

First, there are the contrasting views on how these economies operate and how they respond to various policy tools. Under the best of circumstances policymakers would receive conflicting advice. The ongoing controversy about IMF programs is an indication of this aspect. Second, some of the civil servants entrusted with implementing the policies decided upon by the policymakers may not respond in the required fashion. For example, it is easy to change a tax law; it is much more difficult to make the tax administrators fully implement the change. Third, statistics that are essential for good policymaking are often not available or are available with considerable delays or with sizable errors. Fourth, changes in policy instruments are often neutralized by the reaction of forces outside the control of the policymakers or even of the civil servants. For example, an increase in import duties or in income tax rates may have little effect on revenue if smuggling is easy and tax evasion is rampant. Fifth, authorities often find unacceptable, for various reasons, policies that may be seen as desirable by economists. Considering all these reasons, one should expect different fiscal responses to exogenous shocks in *LDC*s as compared to industrial countries. There is also the complication that exogenous shocks generate not just fiscal imbalances but also external imbalances, which may not be easily financeable. The policymakers often find themselves in situations where they have to coordinate conflicting objectives concerning internal and external imbalances.

The countries that, in the 1970's, were faced with rising public revenues due to higher export prices generally reacted in three different ways. A first (and very small) group considered the increase as a temporary windfall which would affect only marginally the permanent income of the country and of the government. These countries used the additional revenue to pay off foreign debt or to accumulate foreign assets (in the form of foreign exchange or in real assets). These assets could be liquidated in future years when foreign earnings declined in order to maintain the level of domestic spending on some, hopefully permanent, trend. This behavior is an application of the permanent income hypothesis of consumption to the government.

A second, and larger, group engaged in capital accumulation at home by expanding the level of public investment. Provided that the investment has as high a rate of return as what the country could have received from foreign assets, that the "additional" investment spending is limited to the windfall income, and that this spending can be phased out when the windfall income begins to disappear, this policy response can also be considered as a good one. However, experience indicates that often the requirements mentioned above were not met. Investment was often not as productive as it could have been having been distorted by poor management and by political considerations; it was often too large; and it was too rigid to be phased out when needed. These countries faced difficulties when the windfall disappeared and foreign financing dried up. These changes would have required a quick reduction of the investment expenditure.

A third and largest group increased public spending by increasing public employment, increasing the size of transfers, increasing investment, and so on. In this particular situation, when the decline in foreign earn-

ings inevitably came, the countries were tied to patterns and levels of spending that were very difficult to change. As long as foreign loans were available, the countries used this source to maintain the level of spending that could no longer be maintained with ordinary revenue. This reaction postponed the problem and in many cases made it worse by leaving the countries with huge foreign debts. When the crisis came, and the countries found that they had to adjust, as financing was no longer available, the consequences were very serious.

Shocks that reduce public sector revenue are even more difficult to deal with. In this case, the countries are often unable to make up the revenue losses in the short run. The losses of foreign trade taxes could in theory be compensated by increasing income taxes or taxing domestically produced products. But income taxes take a very long time to introduce and collect, and their scope in *LDC*s is limited. For this reason, countries have often been forced to rely on inferior revenue sources such as inflationary finance, regressive excises, or the building up of arrears.

IV. Conclusions

Unlike the industrial countries, where the government has much greater control over revenue sources, where revenues are rarely tied to the foreign sector, and where there is always the option of selling bonds domestically to generate additional domestic revenue for the public sector in a noninflationary way, in the developing countries the degree of freedom in the policy area is much more restricted for some of the reasons indicated. Another reason is that the generation of domestic noninflationary and nontax sources of revenues is extremely limited. Therefore, in the absence of foreign borrowing, and once the possibility of financing spending through the building up of arrears has been exhausted, there is a limit to the amount of public spending (expressed in real terms or as a share of gross national product) that the government will be able to maintain. This limit is not a rigid one but it exists all the same (see my 1986b paper). Attempting to exceed that limit will bring inflation as the government will have to finance the additional spending through money creation. This channel itself has a limit and inflationary finance may reduce the real value of tax revenue (see my 1978 paper). That absolute limit on real government spending falls when an exogenous shock reduces tax revenue; it falls even more when foreign borrowing is constrained by the unwillingness of the commercial banks to lend to the country. Of course, within the budget itself, to the extent that the servicing of the foreign public debt increases, other expenditures have to be limited even more.

Thus, often the only realistic alternative that these countries have is to reduce public spending. As it is often politically difficult to reduce current spending in the short run, the adjustment pressure is often shifted to capital spending. This is normally seen as an undesirable type of adjustment, although if what is eliminated is unproductive investment projects, it may not be as undesirable as it is often believed.

REFERENCES

Tabellini, Guido, "The Reaction of Fiscal Policies to the 1979 Shock in Selected Developing Countries: Theory and Facts," mimeo., 1985.

Tanzi, Vito, "Inflation, Real Tax Revenue, and the Case for Inflationary Finance: Theory with an Application to Argentina," *IMF Staff Papers*, September 1978, *25*, 417–51.

______, (1986a) "Quantitative Characteristics of the Tax Systems of Developing Countries," in David Newbery and Nicholas Stern, eds., *Modern Tax Theory for Developing Countries*, Washington: International Bank for Reconstruction and Development, forthcoming 1986.

______, (1986b) "External Versus Internal Debt as a Means of Financing Fiscal Deficits in Developing Countries," in Bernard Herber, ed., *Public Finance and Public Debt*, Proceedings of the 40th Congress of the International Institute of Public Finance, forthcoming 1986.

The Effect of the Union Wage Differential on Management Opposition and Union Organizing Success

By RICHARD B. FREEMAN*

What is the effect of the union wage differential on union organizational success? It is common in economic models of unionization to assume that greater differentials enhance the probability that unions will organize a group of workers. After all, won't workers want to join a union the greater are the potential economic benefits from joining? Following this line of argument, some studies of union wage differentials make unionization an endogenous, presumably positive, function of the potential differential.

Here I argue that the conventional view that large union wage differentials increase organization is incorrect for the United States today. It is incorrect because organization is the joint decision of workers and management, not a worker's decision. While potential wage increases are a plus to workers, they reduce profits and thus are a minus to management. In an institutional setting in which management allocates significant resources to convince workers to vote against unions, it is erroneous to analyze the effect of the union wage differential on organizing solely from the workers side, just as it is erroneous to analyze any economic outcome solely in terms of one blade of the market scissors. My claim is threefold.

1) Current institutional facts indicate that, despite the secret ballot election procedure in which only workers vote on whether to organize, the decision to unionize in the United States is dependent on management as well as workers.

2) While in the most general model of organization, the dependence of organizing success on management as well as labor makes the impact of potential wage differentials indeterminate, more structured models suggest that the magnitude of the potential differential will increase management opposition more than it will increase worker desires for organization, causing an *inverse relation* between the differential and union success.

3) Extant empirical evidence for the recent decline in union organization suggests that as much as one-quarter of the decline in the proportion organized through NLRB elections may be attributed to the increased union wage premium of the 1970's and its adverse effects on firm profitability, which raised management opposition.

If my argument is correct, the observed willingness of unions to give concessions to companies in the union wage impact in the 1980's ought to reduce opposition and improve organizational success, at least up to some point.

I. Institutional Facts

Organization of workers through the NLRB procedure currently involves a lengthy confrontation between two organized parties, the workers and their proposed union representative, and management, often abetted by outside union-management consultants. The process is typically long (2 months between the filing of a petition and an election), with numerous possibility for delays and pitfalls.[1]

In most cases management takes an active role opposing organizing, hiring consultants in upwards of 70 percent of campaigns and

†*Discussants*: Charles Craypo, University of Notre Dame; Michael Reich, University of California-Berkeley.

*Harvard University and National Bureau of Economic Research, Inc., 1050 Massachusetts Avenue, Cambridge, MA 02138.

[1]See NLRB, *Annual Statistics*, and Myron Roomkin and Richard Block (1981).

TABLE 1—UNION SUCCESS RATE IN NLRB REPETITION ELECTIONS, BY ROLE OF SUPERVISORS IN THE CAMPAIGN

Role of Supervisors in the Campaign (percent of cases)	Union Success Rate (percent)
None (6)	100
Some (8)	70
Moderate (18)	57
Sizeable (16)	20
Extreme (51)	33

Source: AFL-CIO Organizing Survey, April 1984, Appendix, p. 37.

often breaking the law by firing union activists.[2] (There are $1\frac{1}{2}$ illegal firings per NLRB election, according to NLRB data.)[3] Management campaign tactics range from personal letters, in-plant meetings, supervisors' discussions, and a wide variety of propaganda in the form of leaflets, posters, and so on. As a crude indication of the potential effectiveness of such tactics, the evidence from the *AFL-CIO Organizing Survey* shows an extremely strong relation between the role of supervisors of campaigns and union success (see Table 1).

Given the importance of management activities in union success we must analyse how the union wage differential affects those activities as well as worker desires for unions to see how the differential affects outcomes of NLRB elections.

II. The Wage Differential, Labor and Management Organizing Effort, and Organizing Success

Consider first the most broad (and least informative) model of union organization in which both management and labor affect the outcome. The vote for unionization (V) is taken to be a function of "objective" circumstances (X) and of the resources spent by labor (R_L) and management (R_M) on the organizing campaign:

$$(1) \qquad V = V(R_L, R_M, X).$$

[2]AFL-CIO Department of Organization and Field Services (1984).

[3]Calculated from NLRB data for 1980. See Paul Weiler (1983).

The resources allocated to the organization drive will depend on exogenous factors specific to management R_M and labor R_L, and the logarithmic (or percentage) wage differential (W_μ):

$$(2) \quad R_M = g(W_\mu, X); \quad R_L = h(W_\mu, X).$$

From (1) and (2) the effect of the wage differential (here taken as exogenous, though in a more complete model it will depend on elasticities of labor demand and other factors) on organization is ambiguous. It depends on the amount of resources it induces both parties to invest in the campaign and the effectiveness of those resources.

$$(3) \quad dV/dW_\mu = V_1\, dR_M/dW_\mu + V_2\, dR_L/dW_\mu,$$

where $V_1 < 0$ and $V_2 > 0$.

At this level of generality all that one can say is that to make unionization a positive function of potential wage gains is erroneous, because it ignores the effect of those gains on profitability and thus on management resources devoted to defeating unions in an organizing campaign.

Under seemingly plausible assumptions one can go further and show that the wage differential is more likely to deter than to increase organization. Assume that management and labor resources have the same effect on outcomes (when $R_M = R_L, V_1 = V_2$.) Then the standard monopoly analysis of union wage gains suggests that management will increase its organizing resources more than will a union as the wage differential rises. Figure 1 depicts the essential argument in terms of a standard labor demand analysis of the welfare effects of union monopoly wage gains. Here W, L are wages and employment in the absence of unionism; W' and L' are wages and employment in the presence of union wage premium $W\mu$.

The wage differential $W' - W$ transfers $(W' - W)L'$ dollars to labor but costs management $(W' - W)L' + \frac{1}{2}(W' - W)(L' - L)$, where the latter term is the welfare triangle loss from the higher wage. With a given union wage premium $W_\mu \approx (W' - W)/W$

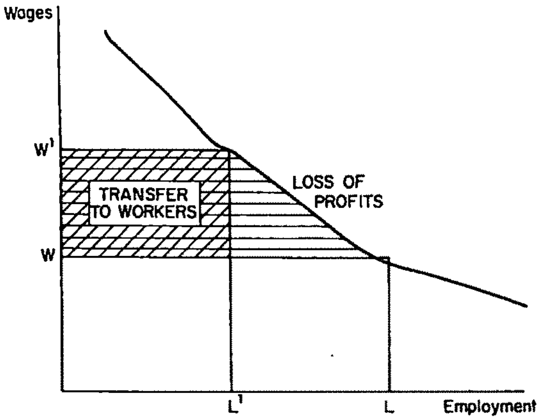


FIGURE 1. THE EFFECT OF THE UNION WAGE PREMIUM ON MONEY GAINS TO WORKERS AND LOSS OF PROFITS TO EMPLOYERS

and an elasticity of demand for labor of η, the welfare loss is $\frac{1}{2}\eta[W\mu]^2 WL$.

Labor will be willing to spend the rectangle to organize. Management will be willing to spend the rectangle plus the triangle. Assume that management and labor do, indeed, spend the maximum amounts possible. Then

$$R_M - R_L = \tfrac{1}{2}(W_\mu)^2\eta(WL). \tag{4}$$

Differentiating we see that $d(R_M - R_L)/dW_\mu = \eta W_\mu(WL) > 0$. Management will devote greater resources than labor in an organizing campaign and will increase those resources more as the wage differential increases.

The model given in equations (1)–(3) and the figure is simplistic. It can be developed in various ways (more complex reaction functions; consideration of unions and workers as separate groups; different expectations of W_μ; and so forth). The point is simply that unless one believes that unions efficiently extract "rent" from firms along the lines of efficient contract models, it is theoretically reasonable to expect the union wage differential to generate more management opposition than worker and union support in organizing drives.

III. Empirical Evidence

I present two types of evidence on the actual impact of the union wage differential on unionization: a time-series analysis of 1950–80 changes in the union wage premium, management unfair labor practices, and the number of workers organized through NLRB elections; and a 1965–80 pooled cross-industry time-series analysis of the effect of unfair labor practices on workers organized across industries.

Table 2 records the results of the time-series analysis. Equation (1) gives the regression coefficients for the determination of my indicator of management opposition—unfair labor practices (8a3 violations) per worker. Equation (2) gives the regression results of the effect of those practices on numbers of workers won in elections per employee while equation (3) shows the "reduced-form" effect of the wage differential on workers won per election. The control variables include time and three indicators of the general state of the economy.

There are three findings. First, consistent with my argument that the union wage differential increases managerial opposition the coefficient on W_μ in the unfair labor practice equation is positive and significant. Second, the unfair labor practices variable has a marked negative effect on workers won in equation (2) while the wage premium has a negative effect in equation (3), supporting my argument that the union wage advantage adversely affected organization. Third, however, as the significant coefficients on the time trend variable indicate, my analysis falls short of a complete explanation of the pattern of organization in the period.

To see how much of the observed change the key variables explain, I have multiplied the regression coefficients by the change in the relevant explanatory factor and divided this by the observed change in the explanatory variable. For 1970–80, 47 percent of the increase in unfair management practices are attributed to the rising wage premium; one-half of the decline in workers won per employee is attributed to the rise in unfair practices; and two-thirds to the rise in the wage premium, taken by itself. Over the longer haul, however, the significance of the wage differential is, it should be noted, smaller.

All told, while the story is far from complete, the data suggest that in the 1970's the

TABLE 2—ESTIMATED EFFECT OF UNION WAGE PREMIUM ON UNFAIR LABOR PRACTICES ON ORGANIZING SUCCESS, 1950–80

Independent Variable	Dependent Variable: log(Unfair Labor Practices [8a3] per Worker in NLRB Election (1)	log(Number of Workers Organized/ Employment) (2)	(3)
Constant	−4.3	−6.62	−5.80
Time	.15	−.08	−.14
	(.04)	(.04)	(.04)
Real *GNP*	−.002	−.000	.003
	(.001)	(.001)	(.001)
Rate of Inflation	−1.35	01.83	−2.07
	(1.43)	(1.52)	(1.71)
Unemployment Rate	.02	−.05	.02
	(.04)	(.04)	(.04)
W_μ (percentage wage differential)	2.45		−2.98
	(1.04)		(1.16)
log(unfair labor practices per worker)		−.46	
		(.20)	
R^2	.96	.92	.91

Source: Unfair practices, workers organized in NLRB elections; NLRB *Annual Report*, various editions; wage differential, from George Johnson (1981); rate of inflation, unemployment rate, and real *GNP* from *Economic Report of the President*; 1982.

sizeable increase in the union wage differential augmented management opposition and contributed to the decline in union organizational success.

Economists are, rightly, suspicious of the results of short time-series analysis, like those in Table 2, which often vary depending on model specification and years covered. Accordingly, I have also estimated the effect of unfair labor practices on members won by unions using a pooled *cross-section industry* file for the period 1965–80 over which industry data was available. The advantage of this data file is that it permits two separate types of analyses: first, a comparison of patterns of unfair practices and members won *across industries*; second, an extremely strong test of the effect of the factors within industry-year cells. The disadvantage is that we lack information on union wage differentials by industry and thus can only examine effect of management opposition on union success or failure in elections, and cannot estimate the effect of wage differentials on management unfair labor practices.

TABLE 3—ESTIMATES OF THE EFFECTS OF UNFAIR LABOR PRACTICES ON ORGANIZING SUCCESS IN A POOLED INDUSTRY TIME-SERIES MODEL 1965–80

Independent Variable	Dependent Variable: Log (Number of Workers Won/Employment) (1)	(2)
log (unfair practices per election)	−.36	−.62
	(.07)	(.06)
Control Variables		
Producer price	X	X
Profits	X	X
Year Dummies	X	X
Industry Dummies		X
Wages	X	X
R^2	.15	.70

Source: Workers won, unfair labor practices, from NLRB *Annual Statistics*. All other data from U.S. Bureau of Labor Statistics.
Note: Number of Observations: 684 with some industry-year cells missing in early years.

Table 3 shows the results of least squares estimates of the effect of unfair labor practice on number of workers won by unions. Consistent with the results in Table 2, these figures show a sizeable and significant impact of management opposition—as measured by unfair practices—on organizing success. The greater the number of illegal acts by management the less likely are unions to win members in one industry compared to another (col. 1) or to an industry compared to itself over time (col. 2).

IV. Conclusion

While theory does not tell us whether union wage premiums raise or lower organizing success, it does tell us that with current U.S. institutions, the effect of the premium depends on what it does to both labor and management behavior and gives some reason for thinking that the higher premium may raise managerial opposition more than it raises worker desires for unionization. The empirical analysis attributes part of the union problem in organizing in the 1970's to managerial opposition resulting from high union wage premiums. Hopefully, the analysis and finding will stimulate further work on

organization as the result of the behavior of both management and labor, in contrast to existing focus on workers alone.

REFERENCES

Johnson, George, "Changes over Time in the Union/Nonunion Wage Differential in the U.S.," mimeo., University of Michigan, 1981.

Roomkin, Myron and Block, Richard, "Case Processing Time and the Outcome of Representative Election: Some Empirical Evidence," *University of Illinois Law Review*, 1981, *5*, 75–97.

Weiler, Paul, "Promises to Keep: Securing Workers' Rights to Self-Organize Under the NLRA," *Harvard Law Review*, June 1983, *96*, 1769–827.

AFL-CIO Department of Organization and Field Services, *AFL-CIO Organizing Survey*, Washington, April 1984.

Council of Economic Advisors, *Economic Report of the President*, Washington, 1982.

National Labor Relations Board, *Annual Statistics*, Washington: USGPO, various years.

Union-Nonunion Earnings Differentials and the Decline of Private-Sector Unionism

By RICHARD EDWARDS AND PAUL SWAIM*

Recent years have been difficult ones for the American labor movement. Especially during the past half-decade, the economic and political environment for unions has become increasingly hostile—dominated by a growing anti-union sentiment in management; the adverse effects of industrial restructuring, import competition, deregulation, and high unemployment; and the tightening constraints of a labor law and NLRB enforcement mechanism that have become markedly less supportive of unions. These and other external changes, as well as some continuing internal weaknesses, have thrust unions into a period of declining union membership, eroding bargaining strength, repeated contract concessions, and what at least some observers have perceived as the beginning of a "new era" in industrial relations (Thomas Kochan and Michael Piore, 1985; Edwards and Michael Podgursky, 1986).

In the specific area of wage setting, recent and continuing concession bargaining by unions has attracted the most attention. Union retreats in steel, autos, and transportation have been highly publicized, but several recent studies have suggested that the decline in union bargaining strength has extended as well to construction, retail foodstores, and other industries not immediately affected by such pressures as import competition or deregulation. A large and unprecedented fraction of settlements now involve wage freezes or reductions. And, if current union members face bleak settlements, the proliferation of two-tier wage provisions may presage even more severe cuts for future employees (Charles Craypo, 1981; Daniel Mitchell, 1985).

*Professor and Assistant Professor of Economics, respectively, University of Massachusetts, Amherst, MA 01003. This paper is part of a joint project undertaken with Michael Podgursky. We thank Marc Kitchel for excellent research assistance.

Despite these developments, there remains the question of whether bargaining under these conditions has in fact resulted in a substantial compression of union-nonunion earnings differentials. The union-nonunion wage effect has typically been estimated by cross-section studies using micro data on union and nonunion workers (Richard Freeman and James Medoff, 1981, Table 1). But until recently it has not been possible to replicate these studies for the 1980's, since collection of union membership data in the *Current Population Survey* (*CPS*) was temporarily suspended.

Drawing upon recently released *CPS* data for 1984, we find that the union relative wage gap in the private sector remains substantial and does not seem to have narrowed since 1979. Indeed, in contrast to what perhaps has been commonly assumed, union and nonunion wages may actually have drifted farther apart. While the causes of this outcome remain to be studied, we conclude that the most striking aspect of recent experience is not a rapid compression of union earnings premiums, but rather their general persistence. In effect, labor's changing circumstances seem to be reflected in a quantity rather than a price adjustment: the union wage differential has not changed, and instead there has been a rapid substitution of nonunion for union workers.

I. Model and Results

In order to determine the effect of unionization on earnings we estimated the following reduced-form earnings equation:

$$\ln(W_i) = B_1 X_i + B_2 U_i + e_i. \tag{1}$$

The dependent variable is the natural log of hourly earnings for the ith worker. The term X_i is a set of independent variables including demographic variables, education, region, oc-

cupation, and industry dummies; these variables capture variation in human capital investments and other worker and job characteristics influencing earnings. The U is a dummy variable taking the value one if the worker is covered by a collective bargaining agreement at his or her current job. Thus, the estimated coefficient of this variable measures the natural log of the ratio of union to nonunion earnings, which is approximately the proportionate union-nonunion earnings gap (the exact proportionate differential is $\exp(B_2)-1$). We use a union coverage variable rather than a union membership dummy since nonunion workers covered by collective bargaining agreements (some 2–3 percent of the labor force) by law receive identical benefits as union members. Hence, a union membership dummy may yield a biased estimate of the union wage effect. Finally, we assume that unmeasured variation in remunerative worker and job characteristics is uncorrelated with our independent variables and is imbedded in the mean-zero independent and identically distributed residual e_i.

Equation (1) was estimated on a large sample of nonfarm private wage and salary workers from the May and June, 1984 *CPS*. In order to capture industry interaction effects, we stratified the sample by broad industry groups and estimated equation (1) within each group. For comparison we estimated the same models on a smaller sample of workers from the May 1979 *CPS Pension Survey*, which includes matched June earnings records (Wesley Mellow, 1983). In both samples we excluded managers, professional workers, and private household workers in order to focus on a somewhat more homogeneous and organized segment of the labor force. F-tests allowed us to reject the hypothesis that the earnings equations were identical across these broad industry groups. Race, sex, and Hispanic ethnicity union interactions were insignificant in the stratified regressions in both years.

Ordinary least squares estimates of the union coverage coefficient (B_2) and related statistics are reported in Table 1. The coefficient of union coverage took its expected positive coefficient and was highly significant in all sectors in both 1979 and 1984. The union coefficient differed significantly across sectors, and in 1984 it was lowest in Manufacturing and highest in Construction and Transportation, Communication, and Public Utilities. This cross-section pattern may reflect a more elastic demand for union labor in Manufacturing since offshore production, plant relocation, or imports do not provide as ready substitutes for union labor in the latter industries as they do in Manufacturing.

Our focus, however, is not on cross-section patterns, but on changes over time—and here the results are rather surprising. The unionization rate in our sample declined from 27.8 to 19.0 percent over these five years, reflecting the sharp medium-term decline reported by the BLS and continuing the longer-term trend noted by many researchers

TABLE 1—ESTIMATED UNION-NONUNION EARNINGS DIFFERENTIALS AND UNIONIZATION RATES: MAY–JUNE 1979, AND MAY–JUNE 1984[a]
(Dependent Variable = Natural Log of Hourly Earnings)

	May–June, 1979		May–June, 1984		
Industry	$\bar{U}$	$\hat{B}$	$\bar{U}$	$\hat{B}$	$\hat{B}$
Mining, Forestry and Fisheries	.470	.222[c] (.049)	.204	.218[c] (.062)	−.004 (.079)
Construction	.375	.369[c] (.027)	.274	.436[c] (.024)	.067 (.036)
Manufacturing	.440	.130[c] (.011)	.320	.150[c] (.011)	.020 (.016)
Trans., Comm., and Public Utilities	.615	.229[c] (.026)	.462	.300[c] (.023)	.069[b] (.035)
Trade and Services	.112	.202[c] (.015)	.088	.217[c] (.014)	.015 (.021)

Sources: May 1979 *CPS Pension Survey* and 1984 Earnings File. Microdata tapes available from the Bureau of Labor Statistics.

[a]Means and *OLS* estimates of the coefficient of union contract coverage dummy variable were obtained by estimating equation (1) within each of the industrial groups shown above. Within each industry group, managers, professionals, and private household workers were excluded. In addition to union coverage, the regression included controls for education, race, Hispanic ethnicity, sex, years of labor market experience, household head, veteran, SMSA residence, region, part-time work, and 7 occupation dummy variables. Industry dummies varied with the sample, ranging from 1 in Mining, Forestry, and Fisheries, to 16 in manufacturing. A complete set of regression coefficients and related statistics are in a separate appendix available from the authors.

[b]Significant at a .05 level of confidence.

[c]Significant at a .01 level of confidence.

(Larry Adams, 1985; Freeman and Medoff, 1984, ch. 15). Sharp declines also occurred within each of our broad industry groups, ranging from approximately 3 percentage points in services to over 25 percentage points in Mining, Forestry, and Fisheries.

In spite of very sharp declines in the unionization rate, the union-nonunion earnings differential remained relatively stable or widened within each of these broad industrial sectors. The last column of the table shows that, with the exception of Mining, Forestry, and Fisheries, the estimated union-nonunion differential widened in every industry group examined. The widening is statistically significant (and then only at a 5 percent level of confidence) only in Transportation, Communication, and Public Utilities.

The estimates in Table 1 constrain the union wage effect to be the same across industries within these rather broad industrial groups. We also examined possible industry-union interaction effects within several of these broad industrial groups. In Transportation, Communications, and Public Utilities, we tested whether the union-nonunion wage gap in recently deregulated industries exhibited a different trend. The only significant interaction was trucking services, where the union-nonunion gap dropped from .328 to .308, a statistically insignificant decline. Union-industry interactions in the earnings equation for manufacturing showed a widening gap in nine industries and narrowing gap in nine others. In no industry, however, was the change significant at a 5 percent level of confidence. In four cases the change was significant at a 10 percent level of confidence: three with a widening gap (Food and Tobacco, Basic and Fabricated Metal, and Electrical Machinery and Equipment); and one with a compressed gap (Rubber and Plastic Products). The 1979 employment-weighted sum of the changes was +1.2 percentage points, approximately equal to the change reported in Table 1.

II. Discussion

These findings are surprising given the considerable attention that has focused on concession bargaining by unions. Before attempting to explain them, we note that our findings are consistent with more aggregated wage-trend data published by the Bureau of Labor Statistics. The Wage and Salary Employment Cost Index indicates that between 1979 and 1984 union wages rose 5.1 percent more than nonunion wages in manufacturing and 5.6 percent more in nonmanufacturing industries (U.S. Department of Labor, 1985, p. 38). Nonetheless, we view the results in Table 1 as preliminary since the matched May–June 1979 sample may not be fully comparable with the 1984 sample, which simply pools May and June *CPS* records (difficulties in matching records could have resulted in nonrandom attrition from the sample). We are currently constructing data files for 1978 and 1979 that avoid this potential problem.

Perhaps the most straightforward interpretation of these findings would focus on price-quantity adjustments. It could be argued that, faced with increasingly elastic demand for their members' services, most unions have evidently chosen to maintain established wage levels; the result has been dramatic reductions in employment. This interpretation, of course, is at odds with journalistic accounts emphasizing a perceived willingness of many union negotiators to trade substantial wage and benefit concessions for enhanced job security. It would similarly confound Colin Lawrence and Robert Lawrence's provocative 1985 analysis of recent wage developments in manufacturing, since their analysis relies heavily on the notion that union wage demands vary inversely with the elasticity of demand. But, in this interpretation, our results would indicate that unions assign a lower priority to maintaining membership levels than is commonly believed.

There are, however, several reasons to be cautious about this wage inflexibility interpretation. First, in the Wage and Salary Employment Cost Index mentioned above, nonunion earnings growth exceeded union earnings growth in manufacturing industries in 1983 and 1984, and in nonmanufacturing industries in 1984. Thus, it may be that concession bargaining has only very recently

become widespread enough to be reflected in broad wage averages. In the same vein, as was already mentioned, two-tier wage scales have sometimes been implemented that partially shield existing employees from the full brunt of the compensation cuts that have been negotiated. If these provisions remain in effect, the impact on earnings will continue to grow as the current work force is progressively replaced by new workers.

Second, while our list of control variables is extensive and includes general labor market experience, we were not able to control for a possible rise in the average seniority of unionized workers, since data on employer-specific job tenure was not collected in the 1984 *CPS* surveys. Seniority plays an important role in layoff and recall in most union contracts, so it is highly likely that workers who have weathered the shakeout in the union sector have greater seniority relative to an average nonunion worker in 1984 as compared to 1979. Thus, the coefficients for union coverage in the 1984 earnings regressions could be biased upward relative to the 1979 coefficients and the extent of union wage flexibility correspondingly understated.

The May 1979 *CPS*, which included information on employer-specific seniority, showed, not surprisingly, a strong correlation between seniority and age. Indirect evidence on the magnitude of effect of changes in seniority on estimated union wage effects can thus be gleaned from an examination of union-nonunion age gaps. In 1979, union workers were already older than nonunion workers, yet these average gaps widened appreciably between 1979 and 1984 in all sectors except trade and service. If we assume that the same five-year changes apply to comparisons of average seniority levels between union and nonunion workers, then the product of these estimated increases in average seniority gaps with estimates of the marginal earnings effect of more seniority provides an indication of the extent of upward bias in the union coverage coefficients. The results of these calculations (described in a statistical appendix available upon request), indicate that plausible shifts in average seniority levels were appreciable, but probably too small to have obscured a substantial compression of union sector wages.

The apparent predominance of quantity adjustments over price adjustments revealed in Table 1 may also reflect a second compositional effect that further obscures the extent of union wage flexibility. Just as higher seniority members were probably more successful in negotiating the rapid decline in union membership within each bargaining unit, differential survival probabilities across bargaining units may also have played a role in maintaining the union wage differentials inherited from the 1970's. One source of the rapid decline in union membership has been the loss of locals representing workers at establishments that were either closed or reorganized to operate nonunion. This attrition may have been most pronounced in establishments where unions already failed to negotiate compensation levels commensurate to those achieved by other union locals. A "survival of the fittest" effect could then have resulted in stable or rising average union-nonunion earnings differentials even though many surviving unions made substantial concessions: the "give-backs" negotiated by surviving unions being offset by the disproportionate elimination of the weakest union contracts from the universe of current contracts.

Although the *CPS* data provide no information with which to test the hypothesis just offered, it is consistent with the higher compensation levels typical of larger employers and with what anecdotal evidence suggests is a more rapid de-unionization of small firms. In some manufacturing industries, for example, where the large, core employers remain unionized, the deunionization of the periphery has been encouraged by the expansion of outsourcing to small, nonunion firms. Similarly, in trucking and construction, large unionized core employers have themselves acquired small nonunion peripheral firms in order to expand into new markets—a phenomenon termed "double-breasting" by unions. Both of these practices particularly disadvantage small, unionized firms that find themselves in direct competition with these nonunion producers. The higher mortality

rates of small establishments (David Birch, 1979) combined with the low and falling success rate of unions in representation elections (Freeman and Medoff, 1984, ch. 15) also suggest that union penetration has eroded most rapidly in smaller establishments. Thus, the union-nonunion gap may have widened not because unions continued to obtain relatively higher wage settlements than their nonunion counterparts, but because union control of their jurisdiction is slipping away most rapidly where wages are lowest.

III. Conclusion

Our examination of the changes in unionization rates and earnings differentials between 1979 and 1984 indicates that nonunion workers have very rapidly been substituted for union workers in the face of a stable or slightly widened union-nonunion earnings differential. The failure of the earnings differential to decline is surprising, given the widely reported and apparently substantial erosion of union bargaining power.

Several different interpretations of these results are possible. One could interpret them as suggesting that in fact union bargaining power has not declined. This seems unlikely and inconsistent with the sharp fall in unionization rates and much other evidence. Alternatively, one could argue that our results indicate a lack of wage flexibility in the union sector: while unions have been able to maintain the relative price of their labor, in so doing they have placed the burden of adjustment to their deteriorating circumstances on the quantity response. Such an interpretation might be used to suggest that if unions had been willing to reduce the union wage premium, the resulting quantity adjustment would have been less.

A third explanation would place more stress on the institutional relationships involved to suggest that our results may be masking a somewhat more complicated story. Union jobs may have been lost in marginal and peripheral enterprises due to import penetration, deregulation, and the relative ease with which union workers could be replaced by nonunion workers, resulting in the subsequent retrenchment of the union sector to a pool of older, high-tenure workers in a shrinking core of large establishments. If true, these compositional shifts have apparently more than offset the very poor (by historical standards) wage settlements that a number of unions have been forced to accept. Under these circumstances, there may have been no reasonable concessions that unions could have made that would have stemmed the loss of union jobs.

The last interpretation, while consistent with fragmented and anecdotal data, is obviously speculative and requires confirmation from further survey data. Moreover, the second and third explanations are not mutually exclusive and may be valid in differing degrees for different industries. Finally, the impact of continued concession bargaining on broad measures of union wage effects may become much more pronounced as these compositional effects run their course.

REFERENCES

Adams, Larry T., "Changing Employment Patterns of Organized Workers," *Monthly Labor Review*, February 1985, *108*, 25–31.

Birch, David L., *The Job Generation Process*, Cambridge: MIT Program on Neighborhood and Regional Change, 1979.

Craypo, Charles, "The Decline of Union Bargaining Power," in M. Carter and W. Leahy, eds., *New Directions in Labor Economics and Labor Relations*, Terre Haute: Notre Dame University Press, 1981.

Edwards, Richard and Podgursky, Michael, "The Unraveling Accord: American Unions in Crisis," in R. Edwards et al., eds., *Unions in Crisis and Beyond*, Dover: Auburn House, 1986.

Freeman, Richard and Medoff, James, "The Impact of Collective Bargaining: Illusion or Reality?," in Jack Steiber et al., eds., *U.S. Industrial Relations, 1950–1980: A Critical Assessment*, Madison: Industrial Relations Research Association, 1981, 47–98.

______ **and** ______, *What Do Unions Do?*, New York: Basic Books, 1984.

Kochan, Thomas A. and Piore, Michael J., "U.S.

Rising Union Premiums and the Declining Boundaries Among Noncompeting Groups

By PETER LINNEMAN AND MICHAEL L. WACHTER*

The notion that unions are in decline in the United States is an important theme in current discussions of labor markets. The dimensions of this decline, however, are seldom analytically identified. Instead, the debate has primarily focused on the decline in unionized employment relative to total employment. Although unions' share of total employment has been declining for several decades, this process has quickened over the past decade. Today, the absolute number of unionized workers, as well as the share, is in decline. Other indicators of union strength and health, however, are providing very different information.

After several decades of cyclical but relatively trendless fluctuation, union relative wage premiums have expanded to all-time highs. The magnitude of this increase is impressive, with premiums in some industries increasing by over 50 percent in the last decade. These developments represent a clear break with the past pattern as models that explain cyclical variations in union wage premiums cannot predict either the magnitude of the change or the steady nature of the increase.

Although downward-sloping labor demand curves suggest an obvious link between these two major developments, to date there has been little attempt to connect them. In this paper we analyze union wage premiums, at the one-digit industry level, for 1973 through 1984. Wage premiums over time are then used to explain the variation in union's share of employment by industry and total union employment by industry. We find that these union employment measures are negatively and significantly related to the increases in union wage premiums.

The previous failure to connect the large declines in union employment shares with rises in union wage premiums reflects a traditional view that labor demand functions in the unionized sectors are very inelastic. An important variant of this view—John Stuart Mill's "noncompeting groups"—illustrates the long history of this approach to the subject. Similarly, Richard Freeman and James Medoff (1984) have implicity argued that wage premiums may not have a large impact on union employment because of the increased productivity impact of the exit-voice role played by unions. Their view is related to the hypothesis that if management is satisficing, unions may only cause increased management attention to efficiency.

The potential efficiency effect of an internal labor market, explored by Oliver Williamson et al. (1975), indicates that unions are not a necessary condition for solving the inherent bilateral monopoly problems that exist in that market. Further, an important set of empirical studies by Barry Hirsch and John Addison (1986) and Richard Ruback and Martin Zimmerman (1984) provide evidence that union productivity effects are negligible.

Past research has focused on unionization effects within manufacturing, a sector which, surprisingly, is becoming increasingly unrepresentative of the unionized sector. In analyzing the impacts of relative wages, we find that the critical variation is among the goods and service producing sectors and little can be inferred from examining the manufacturing sector alone.

Instead of matching the decline in unionization with the increase in relative union wages, three alternative explanations for the decline in union employment are developed in the literature. First, the struc-

*Associate Professor of Finance and Professor of Economics, Law, and Management, respectively, University of Pennsylvania, Philadelphia, PA 19104. Research support was provided by the Institute for Law and Economics, University of Pennsylvania. Research assistance was provided by William Carter and Nancy Zurich.

tural shift in the U.S. economy toward service and away from goods-producing sectors is viewed as a primary source of the long-term decline in union membership. This argument holds that the costs of unionizing service workers are large and that the Hicks-Marshall conditions make it unlikely that any resulting unions could achieve large wage premiums. A more activist anti-union attitude among management is cited as a second factor causing the decline of unions. Simply stated, union membership is directly related to the outcome of certification and decertification elections, and unions have been losing more of these elections than in the past. The literature treats these developments as reflecting an exogenous increase in the anti-union sentiment among firms doing business in the United States. Third, the growth in international competition coupled with domestic deregulation is commonly used to explain job losses in previously protected union strongholds.

The existence of a strongly negative employment effect, however, suggests that not only are these explanations incomplete, but also that a component of these three standard explanations for the decline in union employment may be endogenous. That is, although the long-term trend toward a service economy is based on fundamental shifts in demand, the acceleration of this shift since 1973 almost certainly reflects cost factors associated with rising union wage premiums. Similarly, the ability of sectors to compete in what are essentially international markets is clearly affected by relative factor costs *within* the United States. As the relative cost position of the U.S. unionized sector rose over the last decade, the comparative disadvantage of those sectors suffered compared to other domestic sectors with declining relative costs. *Ceteris paribus*, a large change in the relative wage cost of one sector is likely to cause a large shift in comparative advantage. Large and increasing union wage premiums thus have an important impact on which industries and firms can successfully compete in domestic and international markets.

The widening of the union wage differential may also be viewed as influencing the speed of deregulation. To the extent that unions captured the rents from regulation, the value of regulation to policymakers and management is reduced. Firms are more likely to agree to deregulation when they have gained little from the regulation. Similarly, it is possible that policymakers recognized that existing regulation often did not serve the public interest of capturing economies-of-scale in the form of lower prices. Instead higher union wages were translated into higher prices.

Finally, management's attitude toward unions is certainly dependent on the size of union wage premiums, reflecting the perceived costs and benefits of unionization. As relative union wages rose, management opposition to unions would also be expected to grow. In fact, management opposition did grow during the 1970's. The increase in management opposition cannot be explained by regulatory changes. Any change in the union/management regulatory climate has occurred since 1981, as the National Labor Relations Board appointed by the Carter Administration was as sympathetic to unions as any prior Board.

I. Data Base

The data base for this study is the *Current Population Survey* (*CPS*) tapes for the period 1973–84 (excluding 1982, for which unionization data were not available). Hence, the observations are a cross section of individuals for each of the survey years. Since the study requires data on individuals' employment, usual hourly earnings, and union status, the May *CPS* was used for the period 1973–81. For 1983 and 1984, the relevant information was available on a monthly basis, and therefore the number of observations is considerably greater in those years.

For each year, tabulations were made of employment status by industry for union and nonunion workers. In addition, a cross-sectional wage equation was estimated for each year to identify the relative wage structure for comparable union and nonunion workers in the various one-digit industries. The dependent variable in each wage equation was the log of usual hourly earnings.

The independent variables in the wage equation included the standard measures of skill (education and age as a proxy for experience), location variables (geographical region, size of city, etc.), demographic characteristics (race, sex, marital status, etc.), industry, occupation, and union status.

By specifying the wage equation with union and industry interactions, we were able to construct a relative wage structure consisting of union and nonunion wages by one-digit industries for each year. The percentage premiums were calculated using the algorithm described by Robert Halvorsen and Raymond Palmquist (1980). These wage equations were estimated for each year (except 1982) for the period 1973–84. The cross-sectional, time-series relative wage structure was then derived from these equations for the 1973–84 period.

To investigate the impact of wage premiums on employment levels, we restricted the sample to full-time workers. This has the advantage of preventing compositional shifts between full- and part-time workers, not related to relative wages, from biasing our estimates. The limitation is that it probably understates the employment impact of relative wages by ignoring the potential adjustment of substituting part-time for full-time workers.

TABLE 1—CHANGES IN EMPLOYMENT AND WAGE PREMIUMS, 1973–84

	(1)	(2)	(3)	(4)	(5)[a]
Government	0.1	0.6	−0.4	0.1	21.6
Construction	0.2	−1.4	−10.1	0.4	57.7
Mining	0.5	−0.2	15.7	−0.2	62.2
Manufacturing Durables	0.3	−4.4	12.0	−2.0	33.0
Manufacturing Nondurables	−0.7	−2.1	10.9	−1.1	26.9
Transportation	0.9	−0.9	13.0	−0.5	46.4
Wholesale Trade	0.5	−0.2	16.9	−0.1	33.3
Retail Trade	1.5	−0.7	10.0	−0.2	12.9
F.I.R.E.	1.5	−0.1	1.8	−0.0	9.0
Service	2.0	2.4	−2.7	0.1	−1.5
Total Economy	6.8	−6.8	1.8	−3.5	

Notes: Changes in share of total economy employment by sector: Nonunion (col. 1); Union (col. 2)—Change in union wage premium (col. 3)—Change in share of total economy employment due to union premium change (col. 4)—Union wage premium, 1984.

[a]The base group for the premium calculation is a weighted average of sectors with growing employment.

II. The U.S. Industry Employment Structure

In the first two columns of Table 1, we present our tabulations of the change in the percentage (of the *CPS* sample) employment in each unionized and nonunionized sector, respectively, that occurred between 1973 and 1984. These numbers illustrate several important, but largely unappreciated, shifts in the profile of U.S. employment.

First, the decline in employment in the goods-producing industries—durable manufacturing, construction, and mining—is solely located in the union sectors of those industries. Nonunion employment actually increased in each of these sectors. Only in nondurable manufacturing is the employment decline spread over both the union and nonunion sectors. This demonstrates that general sectoral demand shifts were not totally responsible for the overall decline in unionization. Further, the pattern in the goods-producing sectors is a more general phenomenon. Union employment has decreased in every sector except services, finance, and government. On the other hand, nonunion employment has increased in every service-producing sector.

Second, in terms of employment shares, the unions have lost shares in every industry except services and government. In the finance-based industries unions' employment share has been largely stable. These findings indicate that the shift from goods- to service-producing sectors cannot be an important part of the "unions-in-decline" story. The so-called de-industrialization of America appears to be a union-specific phenomenon, at least at the one-digit level. Indeed, the decline in goods-producing employment may be viewed as a result of the heavily unionized nature of employment rather than a shift toward service-producing sectors.

III. Union Wage Premiums

Shifting from employment shares to union wage premiums, it should be recognized that there are several alternative methods for

calculating the union wage premium.[1] The most traditional is the coefficient on the union dummy variable in a cross-section data set. In this form, the premium represents the percentage wage differential, controlling for a set of standard skill and labor market variables, for union workers compared with comparably situated nonunion workers. As such the base group is all nonunion workers. Surprisingly, the premium estimate, based on that methodology, shows little trend between 1973 and 1984. That is, no significant increase in the average aggregate union wage premium is found. The major reason for this aggregate result, however, is that the profile of *union workers* has shifted from the high-premium goods-producing sectors to the low-premium service sectors.

In order to avoid the problem of changing weights in calculating the aggregate premium, we adopt a fixed weighting scheme. Findings with respect to variation in union premiums are independent of the base. However, the magnitude of the premium does depend upon the choice of the base group. Our choice of a base is motivated by an interest in a premium that measures the opportunity wage. Hence, we have weighted the various sectors depending upon their share of employment increases over the period. Sectors with declining employment were given a zero weight. As a consequence, our denominator or base contains all of the nonunion sectors and the union service, finance, and government sectors. Our estimates of union wage premiums for 1984, based on the opportunity wage concepts, are shown in column 5 of Table 1.

Whereas the aggregate series is relatively stable, the one-digit industry premiums calculated from the above fixed base show marked variance in the trend of union premiums by sector. The union wage premium remained largely unchanged from 1973 to 1984 in three sectors—government, finance, and services. Finance and services stand out as having both stable and relatively small premiums. In government, the premium is stable but large (over 20 percent).

Six sectors have had major increases in wage premiums. These sectors are mining, durable manufacturing, nondurable manufacturing, transportation, wholesale trade, and retail trade. In each of these cases the size of the premium increased by more than 10 percentage points. This represents an increase of over 50 percent in four of the six industries. The two exceptions, Mining and transportation, had a large premium prior to 1973, so that the percentage gain in the premium is smaller.

In one sector, Construction, the premium actually declined by 10 percentage points. This result, however, is to some extent an aberration due to the choice of the sample period. The premium in construction is known to have increased significantly in the late 1960's and early 1970's, immediately prior to the sample period.

IV. Relationship Between Premiums and Employment Share

We used the pooled cross-section, time-series data, one observation for each of the eleven years and for each of the one-digit industries in the sample, to estimate a logistic regression relating union employment shares (of total economy employment) to union wage premiums. The methodology and results are fullly described in our discussion paper (1986).

In this study, the union wage premium is treated as exogenous. In the absence of knowledge on the union objective function, one cannot meaningfully specify a simultaneous equation system. While bias may arise from simultaneity between premiums and employment, this is balanced by misspecification problems that would arise in the premium equations.[2]

We find a large and statistically significant (at the 99 percent confidence level) negative impact of union wage premiums on union employment shares. When we allow for different sector impacts, we find that the nega-

[1]For a discussion of union wage premiums, see George Johnson (1984) and H. Gregg Lewis (1986).

[2]Issues in specifying union objective functions are discussed by Ronald Ehrenberg and Joshua Schwarz (1983), Henry Farber (1978), and Daniel Hamermesh (1973).

tive impact of premiums is largest in the goods-producing sectors of the economy and only slightly less negative in service-producing sectors. An insignificant wage premium impact is found in the government sector.

Additionally, we find a negative trend in union employment that is not related to the premium growth. In particular, the year dummies for 1981, 1983, and 1984 have increasingly large and negative coefficients. That is, unionization was abnormally low in these years in all sectors even controlling for the impact of the union wage premiums.

The results of our regression analysis are highlighted in column 4 of Table 1. This column reports the estimated change in each unionized sector's share of total economy employment, which our regression indicates is "attributable" to the change in the sector's wage premium between 1973 and 1984. For durable manufacturing, for example, the actual union share decreased by 4.4 percentage points. Of that total, 2.0 percentage points were "due" to the effect of rising union wage premiums in that sector. More generally, the impact of increases in premium "explains" slightly more than one-half of the union loss of employment share in nondurable manufacturing, transportation, and wholesale trade. In total, our analysis suggests that over half of the 6.8 percentage point decline in the union employment share which occurred between 1973 and 1984 is due to rising union wage premiums during that time period.

V. Conclusion

We have presented evidence of changes in union wage premiums and employment shares over the last decade at the one-digit industry level. This cross-section time-series shows large increases in union wage premiums causing statistically significant decreases in union share of employment for the period 1973–84. In nondurable manufacturing, transportation, and wholesale trade, more than one-half of the employment decline is explained by the increasing premiums. By size (as distinct from share) of impact, the largest negative premium effects are found in mining, transportation, and durable and nondurable manufacturing.

We are tempted to argue that the economic long run is arriving in the 1980's with respect to the existence of the historically large premiums. Boundaries that divide noncompeting groups break down over time as the economic agents find ways to avoid the premiums. To the extent that these types of forces are at work, continuing pressure on union wages or employment can be expected. Structural changes, working through deregulation and international trade, operate with a long lag. In addition, recent court rulings and NLRB decisions concerning such issues as shifting work to nonunion plants or firms have only begun to have an impact.

One unexpected result of our premium estimates is that the upward trend in premiums seems to have continued, at least through 1984. To the extent that concession bargaining is viewed as having started in 1980, there is little visible impact in terms of reversing the premiums. Part of this is to be expected. Many of the concessions simply postponed increases rather than canceled them; the concession to the firm was thus limited to the interest on the deferred payment. In some cases, the concessions only reduced previously agreed upon increases, and the resulting wage growth was still above the rate of wage change in the growing sectors of the economy. In addition, two-tier wage concessions will not affect the premium for some time, but their long-term impact could be quite sizable.

REFERENCES

Ehrenberg, Ronald G. and Schwarz, Joshua L., "Public Sector Labor Markets," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, forthcoming 1986.

Farber, Henry S., "Bargaining Theory, Wage Outcomes, and the Occurrence of Strikes," *American Economic Review*, June 1978, *68*, 262–71.

Freeman, Richard B. and Medoff, James L., *What Do Unions Do*?, New York: Basic Books, 1984.

Halvorsen, Robert and Palmquist, Raymond, "The Interpretation of Dummy Variables in Semilogarithmic Equations," *American Economic Review*, June 1980, *70*, 474–75.

Hamermesh, Daniel S., "Who 'Wins' in Wage

Bargaining?," *Industrial and Labor Relations Review*, July 1973, *26*, 1146–49.

Hirsch, Barry T. and Addison, John T., *Economic Analysis of Labor Unions—New Approaches and Evidence*, Boston: George Allen and Unwin, forthcoming 1986.

Johnson, George E., "Changes over Time in the Union-Nonunion Wage Differential in the United States," in Jean-Jacques Rosa, ed., *The Economics of Trade Unions: New Directions*, Boston: Kluwer-Nijhoff, 1984.

Lewis, H. Gregg, "Union Relative Wage Effects," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, forthcoming 1986.

Linneman, Peter D. and Wachter, Michael L., "Union Wage Premiums and Employment," Discussion Paper, Institute for Law and Economics, University of Pennsylvania, 1986.

Ruback, Richard S. and Zimmerman, Martin B., "Unionization and Profitability: Evidence from the Capital Market," *Journal of Political Economy*, December 1984, *92*, 1134–57.

Williamson, Oliver E., Wachter, Michael L. and Harris, Jeffrey E., "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange," *Bell Journal of Economics*, Spring 1975, *6*, 250–78.

Competition and Cooperation in the Market for Exclusionary Rights

By THOMAS G. KRATTENMAKER AND STEVEN C. SALOP*

Salop and David Scheffman (1983) show that firms profitably can gain market power by conduct that raises their competitors' costs. Raising rivals' costs is a more credible route to market power than is predatory pricing because it is not necessary to cause the rivals to exit, no "deep pocket" is required, and the additional profits are gained immediately.

That paper argues that vertical restraints and contracts with input suppliers can be fertile ground for raising competitors' costs. By contracting with one or more suppliers to exclude rivals, either by dealing with them on discriminatory terms or refusing to deal with them altogether, a firm sometimes can increase its rivals' costs. As a result, it can sometimes gain the power to raise price in the market in which it sells output. This type of contract often can be characterized as the purchase of an *exclusionary right* from the input suppliers. That is, in addition to the purchase of inputs, the "predator" also purchases the right to exclude (some of) its rivals from access to the suppliers' inputs. Exclusionary rights contracts can exist in a variety of forms. At one extreme are agreements that involve only exclusionary rights; no inputs are exchanged at all. For example, it was reported in *Alcoa* that at one time Alcoa purchased exclusionary covenants from power companies from which it did not purchase electricity. The contracts involved only the utilities' promises not to sell electricity to other aluminum companies. Such "naked" exclusionary rights contracts are unusual, of course. Most exclusionary rights are bundled with the sale of inputs. For example, Stroh Beer has alleged that the two major brewers purchase not only advertising time on network sports programs, but also the right to exclude remaining brewers from advertising on those same programs, even though the networks have other advertising time available.

It might be argued that such exclusionary conduct would always fail for two reasons: the excluded rivals would have available effective counterstrategies to prevent their own exclusion; and input suppliers would have no incentives to reduce their sales by excluding some customers. These criticisms imply that raising rivals' costs by contracting with suppliers would not be credible.

It surely is not true that exclusionary strategies to restrict rivals' input purchases will always succeed in raising their costs. For example, where rivals easily can substitute to other equally cost-effective inputs, or where entry into the production of inputs is so easy that the excluded rivals can efficiently produce the input themselves, then cost-raising strategies will fail. Moreover, where competition in the output market would be sufficient to maintain low prices despite the exit or increased costs of the excluded competitors, then no profit-maximizing firm would spend any resources trying to exclude those rivals. These conditions are discussed in detail in our earlier paper (1985).

†*Discussants*: Ronald H. Coase, University of Chicago; Gregory K. Down, Yale University; David Sappington, Bell Communications Research and University of Pennsylvania.

*Professors of Law and Economics, respectively, Georgetown University Law Center, Washington, D. C. 20001.

However, suppose these limitations do not apply. Suppose that input supply and demand are not perfectly elastic. That is, suppose there are entry barriers into input production and rivals' next-best alternatives for input purchases are less cost effective than the foreclosed input suppliers. Moreover, suppose competition in the output market would be insufficient to maintain competitive prices if these rivals' costs were raised, because either supply is upward sloping or tacit price coordination is possible.[1] In this case, *if* the exclusionary rights purchaser would succeed in buying the rights at a reasonable price, it would gain market power in the output market. But, would it be able to buy the rights? Or, would rivals be able to protect themselves from exclusion in this case by offering suppliers more to be *not*-excluded than the predator is offering to exclude them? Would suppliers be willing to sell exclusionary rights to the predator or would they make more by continuing to sell the exclusionary right (and inputs) to the rivals? The remainder of this paper takes up these two interconnected questions.

I. Rivals' Counterstrategies

There is an obvious flaw in the claims that rivals can protect themselves by the counterstrategy of bribing suppliers to *not*-exclude them. If rivals must pay the additional cost of admission to avoid cost increases from exclusion, then the admission fees themselves will serve as the cost-increasing devices. The predator would gain the power to raise price from its rivals' increased costs due to the admission fees rather than by forcing them to buy from more expensive sources of supply. Indeed, the predator would prefer this outcome, since rivals' costs would be increased at lower cost to itself.[2] To use the familiar language of taxicab regulation, it is as if the predator creates scarce medallions that its rivals must purchase in addition to the inputs themselves.

A second flaw in relying on counterstrategies to prevent exclusion arises because the critics apparently misunderstand economic efficiency and the Coase Theorem. The fact that the predator outbids rivals for the purchase of exclusionary rights does not imply that the exclusion is economically efficient. The market for exclusionary rights essentially is a *market for competition*. Unfortunately, even if this market is well-functioning, it will fail to yield the efficient outcome because competition is a classic public good.

Many of the benefits of nonexclusion of rivals are received by third parties who are not involved in the competitive bidding for the exclusionary right—the consumers in the output market. Only if these consumers would join with the excluded rivals in paying the expense of outbidding the predator would inefficient exclusion be prevented. Yet, unless suppliers are selling to a market comprised exclusively of a limited number of large buyers, consumers are unlikely to be sufficiently organized to add significantly to the rivals' bids. Even then, consumers will attempt to free ride on the expenditures of the rivals. Because competition is a public good, society cannot depend on consumers to protect themselves from the adverse effects of exclusion of some sellers by others. Indeed, as Salop et al. (1984) show, competition among producers for exclusionary rights will yield the producer cartel outcome, not a Pareto optimal competitive equilibrium.

Consider the table of profits of the predator and rivals and consumer benefits (see Table 1). Exclusion is socially inefficient: aggregate welfare falls from 525 to 400. Yet rivals would bid only up to 25 to be *not*-excluded, whereas the predator would bid up to 100 to exclude them. Even though suppliers would require a bid premium of 50

[1] This obviously requires entry barriers in the output market. However, if the contracts also exclude new entrants from the input, as is often the case, then the exclusionary contract itself can create the necessary barrier to entry.

[2] This is a simple application of opportunity cost: suppose that competitive suppliers have been charging a rival the competitive input price (equal to marginal cost) of 200. Suppose the purchaser then offers to pay them a consideration of 50 for each unit by which they reduce their sales to the rival. Thus, the suppliers would raise their total unit price to the rival to 250, 50 for the right to buy and 200 for the input.

TABLE 1

	Exclusion	Nonexclusion	Difference
Predator	200	100	100
Rivals	50	75	−25
Suppliers	50	100	−50
Consumers	100	250	−150
Aggregate Welfare	400	525	−125

from the predator to compensate them for their reduced sales, that premium is insufficient for rivals to prevail. The equilibrium of this process maximizes the sum of producer profits (profits of the predator, rivals and suppliers), not consumer or aggregate welfare.

It is not true, of course, that the purchaser's benefits from exclusion always exceed rivals' losses. But, the fact remains that exclusion often can be profitable. A number of factors determine competitors' relative costs and benefits, and thus, the likelihood of successful counterstrategies. These follow because the purchaser of an exclusionary right stands to gain (additional) market power and so is able to bid an amount that reflects those increased profits. Potentially excluded rivals, on the other hand, stand to gain only the more-competitive, nonexclusion price and profit levels, if they are not excluded. If disadvantaged rivals do not exit the industry but only shrink, they benefit from the higher prices on their remaining sales. Thus, as a general matter, the purchaser has more to gain than the rivals have to lose. Only if the industry were able to achieve the collusive outcome, absent the introduction of exclusionary rights, or if potentially excluded rivals were far more efficient than the predator would exclusion not reduce joint profits. Following Salop et al., this analysis has two important implications.

First, exclusion is more likely when the predator is big and the excluded rivals are small. This is because the gains and losses from exclusion depend on the bidders' relative market shares as well as on the price received. This is in striking contrast to predatory pricing, where larger predators bear higher opportunity costs from below-cost pricing.

Second, exclusionary strategies that inflict less total harm on rivals, relative to the benefits to the predator, are more likely to succeed because excluded firms would be willing to bid less to prevent their exclusion. For example, a predator would prefer exclusionary rights that raise rivals' marginal costs relative to increases in their average costs. Because rivals' prices depend on their marginal costs, an exclusionary right that only raises established rivals' fixed costs, without raising marginal costs, will not give the purchaser the power to raise price. Yet, rivals will bid to prevent the injury to themselves. On the other hand, if the strategy only raised rivals' costs at the margin, both the purchaser and the rivals would get the benefits of the resulting price rise on all units, but only bear the higher cost on a few units. Thus, exclusionary rights that raise rivals' marginal costs a lot and raise average costs a little are more likely to succeed. Similarly, strategies to exclude potential entrants are more cost effective because all their costs are variable.

II. Suppliers' Incentives

To prevent counterstrategies, the purchaser must outbid rivals by enough to compensate suppliers for any opportunity cost they bear. Although this limits somewhat the gains to exclusion, it will not change the basic result in most cases. Frequently, suppliers will have alternative outlets for their goods at little loss in revenue, if they reduce their demand by selling an exclusionary right. For example, if Alcoa purchased exclusionary rights for electricity, the suppliers could sell the excess electricity to firms that did not produce aluminum. At the extreme, if the demand by such other firms were perfectly elastic at the pre-exclusion price, then the suppliers would sacrifice no revenue at all; they would simply replace sales to aluminum companies with sales to other customers.

There is, however, a potential "hold-out" problem here. If the purchaser tries to obtain exclusionary rights from a number of suppliers, some suppliers may have the incentive to hold out for a higher price, either because (i) they anticipate being able to extract a higher input price from the rivals, assuming the purchaser succeeds in getting exclusion-

ary rights from others, or (*ii*) they believe that the purchaser can be made to cede more of the monopoly profits it will gain. Of course, if enough suppliers hold out, the predator's exclusion strategy will fail.

A sophisticated supplier will take into account the likelihood of a higher post-exclusion price in calculating its opportunity cost for selling the exclusionary right. This will raise the purchaser's cost and thus reduce, but not eliminate, the number of rights purchased. Applying the analysis of R. Mackay (1984), this can be treated as a monopsony problem. The supply of exclusionary rights rises because suppliers anticipate a higher post-exclusion input price. A purchaser with a significant pre-exclusion market share has a demand curve for exclusionary rights that initially has an upward-sloping portion. This demand curve is steeper than the supply curve. This relationship flows from the purchaser's benefits from the higher post-exclusion output price. The solution to the monopsony problem results in fewer rights being purchased than if suppliers were myopic, but it does not altogether eliminate purchases or undo the strategy.

The second hold-out problem primarily affects the distribution of the monopoly profits between the parties, not the likely success of the strategy. The fact that these "transactions costs" of bargaining can cause some fraction of potential deals to fall through does not disprove the proposition that a large fraction of the deals will succeed enough to raise rivals' costs. In any event, this hold-out problem is unlikely to be serious. Competition among suppliers usually will prevent hold outs. The predator can make a credible commitment to purchase all rights offered at some price or purchase a fixed number of rights from the lowest bidders.[3]

[3]As Mackay points out, credibility is essential because of a third potential hold-out problem. Having purchased some fixed number of rights, the predator would have the incentive to go back into the market and purchase more. Anticipating this opportunism, suppliers may hold out. Mackay also provides the solution: offer suppliers a *most-favored nation* clause that would give them the higher price if more rights are subsequently purchased.

III. Horizontal or Vertical?

A final criticism of our approach is occasionally made by those antitrust professionals who feel the need for simplistic categories. They claim that no additional market power can be created by exclusionary rights contracts. Any competitive problems must be due to an already existing monopoly or cartel. This assertion is incorrect. By purchasing exclusionary rights from more than one supplier, the predator is, in effect, creating the potential for additional horizontal market power.

This response brings forth the following counterargument. Whatever anticompetitive effects are created are due to this horizontal arrangement, not the exclusionary vertical contract. That argument misses the point. In fact, it is the vertical contract that leads to the horizontal effects. Moreover, use of an exclusionary vertical contract can facilitate greater exercise of monopoly power than if the input suppliers simply got together on their own and attempted to raise the prices they charge the excluded rivals. There are two collusive benefits to contracting with a purchaser of rights to exclude its rivals. First, embedding the collusive arrangement in a vertical contract makes it easier to prevent cheating. The purchaser is well situated to monitor cheating and, in the absence of antitrust liability, the contract would be legally enforceable. Second, collusion among suppliers to increase the price they charge the rivals confers a benefit on competitors of those rivals. By bringing such a firm into the arrangement, exclusionary rights contracts allow the beneficiaries to transfer some of their extra profits back to the suppliers, thereby allowing the side payments that may be necessary for successful collusion. For example, this would facilitate input price rises to rivals (even *above* the monopoly price) in which the profits of input suppliers fall a little and the profits of the (potential) exclusionary rights purchaser rise a lot. This analysis is not new, of course. The ability to transfer profits efficiently to input suppliers is the driving force behind much of the economic analysis of vertical integration.

IV. Conclusions

None of this should seem very startling. We have simply shown that a free market in exclusionary rights has little in common with free markets in goods and services. Just as the direct acquisition of rivals by their competitors could sometimes create market power, so can the indirect acquisition that results from controlling rivals' sources of supply.

REFERENCES

Krattenmaker, T. G. and Salop, S. C., "Antitrust Analysis of Anticompetitive Exclusion: Raising Rivals' Costs To Achieve Power Over Price," Georgetown University Law Center, 1985.

Mackay, R., "Mergers for Monopoly: Problems of Expectations and Commitments," FTC Working Paper No. 112, 1984.

Salop, S. C. and Scheffman, D. T., "Raising Rivals' Costs," *American Economic Review Proceedings*, May 1983, *73*, 267–71.

______, ______ **and Schwartz, W.,** "A Bidding Analysis of Special Interest Regulation: Raising Rivals' Costs in a Rent Seeking Society," FTC Working Paper No. 114, 1984; reprinted in B. Yandle and R. Rogowsky, eds., *The Political Economy of Regulation*, forthcoming.

Transforming Merger Policy: The Pound of New Perspectives

By OLIVER E. WILLIAMSON*

That antitrust economics in general and merger policy in particular have undergone significant changes in the past twenty years is undisputed. The reasons for this transformation and the economic consequences, however, are less clear. Although I urge that recent applications of economic reasoning to antitrust have been both significant and beneficial, both claims could be disputed.[1]

The pound of new perspectives under which antitrust has been reshaped during the past decade has been varied (see my 1985 study, pp. 366–70). Given the space constraints, I mainly emphasize merger policy developments to which transaction cost economics arguments have been brought to bear. So as to put this in context, the applied price theory approach and the resulting 1968 *Merger Guidelines* are set out in Sections I and II. The transaction cost economics approach and the 1982 *Merger Guidelines* and related reforms are examined in Sections III and IV.

I. The Applied Price Theory Tradition

What Ronald Coase has referred to as the applied price theory approach to industrial organization is distinguished by two common features: the firm is regarded as a production function, and nonstandard or unfamiliar business practices are presumed to have monopoly purpose and effect. To be sure, there were differences between Harvard and Chicago in this latter respect. Harvard maintained that nonstandard contracting had leverage purposes, while Chicago interpreted the same practices as manifestations of price discrimination. But both led to the same result: "when an economist finds something —a business practice of one sort or another —that he does not understand, he looks for a monopoly explanation" (Coase, 1972, p. 67).

The major differences between Harvard and Chicago concerned entry barriers and the degree of reliance to be placed on structure-performance relations. Entry barriers were featured by Harvard (Joe Bain, 1956) but were held to be conceptually defective (Stigler, 1968) and to result in public policy error (R. H. Bork, 1978) by Chicago. Harvard also placed much more weight on the connections between structure and performance than did Chicago. To be sure, Stigler had once argued that "an industry which does not have a competitive structure will not have competitive behavior" (1952, p. 167); and he subsequently took a very structural view of mergers (see below). But others at Chicago were much more skeptical and presented empirical studies which disputed that concentration was responsible for monopoly margins.

These differences notwithstanding, the view of the firm-as-production-function was featured by both Harvard and Chicago. The "natural" boundaries of the firm being defined by technology, management efforts to expand the reach of the firm beyond technological imperatives arguably had anticompetitive purpose and effect.

Thus Bain held that monopoly is the proximate cause for vertical integration where "clear economies...[involving] a physical or technical" aspect are missing (1968, p. 381). Although Stigler's treatment of vertical integration relied more on prices (the evasion of sales taxes and price controls; the practice of price discrimination) than technology and furthermore included life cycle features

*Gordon B. Tweedy Professor of Economics of Law and Organization, Yale University, New Haven, CT 06520.

[1]George Stigler (1982) expresses grave doubts that antitrust enforcement has been significantly impacted by economics. By contrast, critics of the economic approach to antitrust locate many of the errors and excesses of antitrust enforcement to mistaken applications of economic reasoning. If true previously, why should recent scholarship be regarded more favorably?

(1951), public policy was advised that vertical integration "loses its innocence if there is an appreciable degree of monopoly control at even one stage of the production process" (1955, p. 183). Specifically, a 20 percent firm should be prohibited from acquiring more than a 5 or 10 percent share in any industry from which it buys or to which it sells.

Regimes in which the benefits of vertical integration are narrowly regarded typically treat vertical contracting restrictions even more unfavorably. Thus since vertical contracting restrictions—tie-ins, block booking, customer and territorial restrictions, and the like—are evidently lacking in "physical or technical aspects," monopoly purpose was nominated to fill the gap. The government's argument in *Schwinn* is illustrative. Upon acknowledging that vertical integration sometimes yielded economies, it observed that comparable economies have never been associated with "agreements to maintain resale prices or to impose territorial restrictions of limited duration or outlet restrictions of the type involved here."[2] Accordingly, vertical contracting restraints were regarded "not hospitably in the common law tradition, but inhospitably in the tradition of antitrust."[3]

II. The 1968 *Merger Guidelines*[4]

Although the 1968 *Merger Guidelines* avoided the earlier antitrust mistake of characterizing economies as anticompetitive,[5] very stringent merger standards were proposed and an economies defense was disallowed. The *Guidelines* were informed by the structure-performance approach to industrial organization. Thus the stated purpose of the *Guidelines* was "to preserve and promote market structures conducive to competition" (the theory), there being confidence that knowledge of "market structure generally produces economic predictions that are fully adequate" (the evidence) (p. 6882). The *Guidelines* were mainly expressed in terms of admissible market shares. Horizontal mergers between two 4 percent firms were proscribed in highly concentrated industries; the limit was raised to 5 percent if the four-firm concentration ratio was less than 75 percent. My main interest, however, is in vertical and conglomerate mergers.

A. *Vertical Mergers*

The main antitrust concern with vertical mergers was that these were devices which foreclosed equal access to customers or suppliers. Such monopolistic purpose was ascribed to vertical mergers with market shares even smaller than those recommended by Stigler. The *Guidelines* read as follows:

> [T]he Department will ordinarily challenge a merger or series of mergers between a supplying firm, accounting for approximately 10% or more of the sales of its market, and one or more purchasing firms, accounting *in toto* for approximately 6% or more of the total purchases in that market, unless it clearly appears that there are no significant barriers to entry into the business of the purchasing firm or firms. [Potential economies were not an acceptable defense, in part because] substantial economies...can normally be realized through internal expansion.
>
> [p. 6886–87]

As discussed above, even more severe limits were placed on vertical market restrictions since, under the inhospitality tradition, interfirm contractual restraints lacked even a scintilla of benefit.

B. *Conglomerate Mergers*

Conglomerate acquisitions were held to be anticompetitive if they 1) entailed the removal of a most likely potential entrant from the edge of a market, 2) posed dangers of reciprocal trading, and 3) had entrenchment

[2] Brief for the United States at 50, *United States v. Arnold, Schwinn & Co.*, 388 U.S. 365 (1967).

[3] The quotation is attributed to Donald Turner by Stanley Robinson, *N.Y. State Bar Association, Antitrust Symposium* (1968, p. 29).

[4] U.S. Department of Justice, *Merger Guidelines—1968*, July 9, 1982, para. 4510.

[5] See my 1985 study (pp. 366–67) for elaboration.

effects. Reciprocal buying was described as an "economically unjustified business practice which confers a competitive advantage on the favored firm unrelated to the merits of the product" (p. 6888). Entrenchment concerns were posed by mergers which created size disparities, permitted the combined firms to exercise leverage, and facilitated product differentiation. Firm-as-production-function thinking plainly informed conglomerate policy as well.

III. The Firm as a Governance Structure

The 1968 *Merger Guidelines* were subject to strains of four kinds. For one thing, they violated common sense: mergers were challenged that did not remotely pose anticompetitive concerns. Second, as the importance of economies to economic performance became progressively more evident, the willful sacrifice of economies upon allegations of market power—however slight the magnitude and/or speculative the character—became difficult to square with the public interest. Third, earlier confidence in structure-performance relations was shaken by contrary empirical findings. Finally, the concept of firm-as-governance-structure made progressive headway.

Two sweeping critiques of the firm-as-production-function approach to economic organization appeared in 1972, one by Coase and the other by George Richardson. Both argued that the study of economic organization was inadequately served by the applied price theory tradition and that a much wider range of organizational structures than were admitted by the usual firm or market dichotomy needed to be acknowledged.

Here as elsewhere, it takes a theory to contest a theory. Although Coase had advanced an alternative hypothesis in his classic 1937 article "On the Nature of the Firm"—namely, that transaction cost differences are responsible for the decision to organize transactions in firms rather than markets (or the reverse)—this argument had made little headway against the neoclassical firm-as-production-function construction over the next thirty-five years. There having been no effort to operationalize the transaction cost insight during this interval, the alternative hypothesis had gone nowhere (Coase, 1972, p. 64).

Actually, this somewhat overstates the case. Although transaction cost arguments had made little headway in assessing vertical integration, there was a growing awareness that transaction costs played a central role in understanding market failure. Kenneth Arrow's early puzzlement over the problems that attend markets for information led to the conclusion that a more general formulation of the condition of market failure was needed: "market failure is not absolute; it is better to consider a broader category, that of transaction costs, which in general impede and in particular cases completely block the formation of markets" (1969, p. 48)

Internal organization can thus have transaction cost as well as technological origins. But inasmuch as internal organization experiences transaction costs of its own, the question is what factors are responsible for shifting the balance of these costs one way rather than another.

Posing the issue this way discloses that transaction cost assessments are fundamentally comparative in nature. Such comparisons are facilitated by making the transaction the basic unit of analysis, and by adopting a contracting orientation. What factors are responsible for *differential* transaction cost strains between firm and market organization when each mode of governance is asked to manage an identical set of transactions?

A. *Framework*

A three-part argument is needed. The first entails the realization that contractual difficulties can only be assessed in relation to the behavioral attributes of human actors. Second, transactions themselves need to be decomposed. What are the critical dimensions with respect to which transactions differ? Third, what are the comparative transaction cost disabilities of internal organization? Only the first two are discussed here. The third I have treated elsewhere (1985, ch. 6).

1. *Behavioral Assumptions/Governance.* Transaction cost economics takes issue with

the fiction of comprehensive *ex ante* contracting and with the purported efficacy of court ordering. Cognitive limits (bounded rationality) preclude comprehensive contracting, while incomplete contracts pose hazards if economic agents are opportunistic and dispute resolution by the courts is costly and cumbersome. The upshot is that prior emphases on technology and the courts are pushed into the background and the study of economic organization becomes focused instead on the governance of contractual relations. Economic organization is thus examined with reference to the following imperative: assign transactions (which differ in their attributes) to governance structures (which differ in their costs and competencies) in a discriminating (mainly transaction cost economizing) way. The study of *private ordering* is featured.

2. *Dimensions*. The refutable implications of the transaction cost economics approach to economic organization turn mainly on the effort to dimensionalize transactions. The original argument, which has since been elaborated, is this: recurrent market contracting prospectively poses problems

> if either (1) efficient supply requires investment in special-purpose, long-life equipment, or (2) the winner of the original contract acquires a cost advantage, say by reason of "first mover" advantages (such as unique location or learning, including the acquisition of undisclosed or proprietary technical and managerial procedures and task-specific labor skills).
> [My 1971 paper, p. 116]

More generally, the problem is one of *ex post* small numbers bargaining brought on by a condition of asset specificity, of which four kinds have been identified: site specificity, physical asset specificity, human asset specificity, and dedicated assets (see my 1985 study). The contracting problems posed by asset specificity are due to "potentially appropriable quasi-rents" (Benjamin Klein et al., 1978).

Problems of contract arise only as there is a need to make coordinated adaptations to changing events. The condition of uncertainty is thus also germane, as Steven Wiggins (1985) makes especially evident.

B. *Contractual Ramifications*

Not only does the conception of the firm-as-governance-structure yield a discriminating theory of vertical integration, but the transaction cost approach to economic organization also supports a difference conception of nonstandard forms of contracting and of such economic puzzles as conglomerate organization. Successive studies of vertical contracting restrictions which feature "contracting in its entirety" and the crafting of "credible bilateral commitments" disclose that vertical restraints can and sometimes do economize on transaction costs and encourage investments in least-cost technologies (see my 1985 study, pp. 168–79; 195–205). Severe market power thresholds need to be crossed before contractual restraints are judged to be problematic.

Objections to conglomerate organization because diversification facilitates reciprocity also need to be qualified. Since reciprocal trade (of an appropriate kind) can help to create a mutual credible commitment, more efficient contracting can thereby result. To be sure, only the subset of contracts where trade is supported by investments in transaction specific assets will warrant such an efficiency rationale. Plainly, however, earlier hostility to conglomerates that held that reciprocity was without redeeming economic purpose was overstated.

None of this is to suggest that diversification is an unmixed blessing. It can be and sometimes has been taken to excess. Provided, however, that diversification excesses and "M-form" organizing principles are respected, the conglomerate can be thought of as an internal capital market whereby cash flows from diverse sources are reallocated to high-yield uses. Unanticipated systems consequences, moreover, obtain. Once it was clear that the corporation could manage diversified assets in an effective way, the possibility of takeover by tender offer suggested itself. Managerial discretion was thus restrained by the activation, through takeover, of "the market for corporate control" (Henry Manne, 1965).

IV. The Current *Merger Guidelines*[6]

The main differences between the 1982 and the 1968 *Merger Guidelines* are these: 1) a much broader concept of the market is employed, whence short-run impediments to entry no longer signal a public policy problem as they once did, 2) the benefits of economies are more prominently featured, 3) the vertical merger guidelines expressly reflect transaction cost reasoning, 4) nonstandard forms of organization no longer labor under a presumption of anticompetitiveness, and 5) express provision is made for foreign competition.

Indeed, the 1984 *Merger Guidelines* go even further. An economies defense in cases under review by the Department of Justice is now actively invited. Although this is not without enforcement hazards, the antitrust misconceptions of the 1960's—which led to the suppression, denial, or perverse interpretation of efficiency—have clearly been vanquished.

Not only do the current *Merger Guidelines* make express provision for transaction cost economies, but they acknowledge that the characteristics of investments (especially the condition of asset specificity) are germane to an assessment of economic benefits. Also, vertical integration is now held to be problematic only where the market structure would support strategic behavior, which is also consonant with transaction cost reasoning.

That conglomerate mergers are objectionable because of leverage or reciprocity has been wholly expunged from the recent *Merger Guidelines*—presumably because leverage theory has been discredited and the affirmative purposes of reciprocity are now recognized. Transaction cost reasoning is furthermore reflected in the 1985 *Vertical Restraints Guidelines*, which examine the merits of vertical contracting restraints in a manner very different from the technological orientation/inhospitality predisposition associated with *Schwinn*. A more microanalytic approach to contract in which affirmative transaction cost economizing purposes are admitted is now employed instead. A more permissive set of standards is the result.

[6]The current *Merger Guidelines* are successively set out in U.S. Department of Justice statements issued in 1982, 1984, and 1985.

V. Conclusions

Antitrust works with the economic theories to which it has access. The applied price theory approach to antitrust was much better suited to deal with textbook monopoly problems than it was with complex business structures and nonstandard contracting practices. This is evident from the 1968 *Merger Guidelines*, which use firm-as-production-function reasoning to deal with horizontal, vertical, and conglomerate mergers alike.

A reexamination of vertical integration, conglomerate organization, and vertical contracting restraints under the lens of transaction cost economics discloses that each of these business structures and contracting practices can and often do have redeeming transaction cost features.[7] More permissive antitrust standards are therefore indicated and, properly, have been the result.

Difficult problems of strategic behavior within domestic markets and, even more, in international competition have recently appeared for which additional economic modeling apparatus is needed. Although that too is in progress, the appropriate guidelines are still elusive. Rather than misapply nonstrategic models to complex problems of strategic behavior, the accelerated development of strategic models should be encouraged instead.

[7]Horizontal structures are usefully examined under the lens of transaction cost reasoning as well. The contractual restraints used by ADT in *Grinnell* are an illustration.

REFERENCES

Arrow, Kenneth J., "The Organization of Economic Activity," in *The Analysis and Evaluation of Public Expenditure: The PPB*

System, Washington: USGPO, 1969, 59–73.

Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.

______, *Industrial Organization*, 2d ed., New York: Wiley & Sons, 1968.

Bork, R. H., *The Antitrust Paradox*, New York: Basic Books, 1978.

Coase, Ronald H., "Industrial Organization: A Proposal for Research," in V. R. Fuchs, ed., *Economic Research: Retrospect and Prospect*, Vol. 3: *Policy Issues and Research Opportunities in Industrial Organization*, New York: NBER, 1972, 59–73.

Klein, Benjamin, Crawford, Robert and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Manne, Henry G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, April 1965, *73*, 110–20.

Richardson, G. B., "The Organization of Industry," *Economic Journal*, September 1972, *82*, 883–96.

Stigler, George J., "The Division of Labor is Limited by the Extent of the Market," *Journal of Political Economy*, June 1951, *59*, 185–93.

______, "The Case Against Big Business," *Fortune*, May 1952, *47*, 123 et. seq.

______, "Mergers and Preventive Antitrust Policy," *University of Pennsylvania Law Review*, November 1955, *104*, 176–85.

______, *The Organization of Industry*, Homewood: Richard D. Irwin, 1968.

______, "Economists and the Problem of Monopoly," *American Economic Review Proceedings*, May 1982, *72*, 1–11.

Wiggins, Steven, "The Comparative Advantage of Institutions: Spot Markets, Long Term Contracts, and Firms," unpublished, Texas A&M University, September 1985.

Williamson, Oliver, "The Vertical Integration of Production: Market Failure Considerations," *American Economic Review Proceedings*, May 1971, *61*, 112–23.

______, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

U.S. Department of Justice, *Merger Guidelines —1968*, Commerce Clearing House, Trade Regulation Reports, July 9, 1982. para. 4510.

______, *1982 Merger Guidelines*, reprinted in *California Law Review*, April 1983, *71*, 649–66.

______, *1984 Merger Guidelines*, reprinted in E. Thomas Sullivan and Herbert Ovenkamp, *Antitrust Law: Policy and Procedure*, Suppl., Charlottesville: Michie Co., 1985, 109–30.

______, *Vertical Restraints Guidelines 1985*, reprinted in *Antitrust Law*: *Policy and Procedure*, Suppl., Charlottesville: Michie Co., 1985, 89–108.

Toward a More General Theory of Governmental Structure

By MANCUR OLSON*

A theory of governmental structure begins most naturally with why we need governments. Governments are not needed to perform any functions that markets perform perfectly. Thus a theory of governmental structure naturally begins with market failure. Some economists are uncomfortable with the phrase "market failure" because it is not a sufficient reason for governmental intervention; governments also fail and may perform worse than an imperfect market. This argument, though correct, does not eliminate the need for the concept of market failure. If markets never fail, there is no need for government, so a concept of a market failure is required for an adequate normative theory of government. Since most governments do not restrict themselves to correcting market failures, we cannot explain much of what governments actually do in terms of market failures. Yet we shall see that a theory of governmental structure that begins with a normative analysis of market failure also helps to explain some crucial aspects of reality. Market failure can be quite important, as is shown by the tendency of people throughout history to flee from areas of anarchy to areas with governments, bad as those governments often are. This suggests that the consumers' surplus arising from the elemental services of government is quite large.

Two well-known explanations of market failure can be generalized to cover all types of market failure. When large numbers rule out Coasian bargains, there are two conditions, each of which is sufficient but not necessary for market failure: nonrivalness in consumption and the infeasability of excluding nonpurchasers. Though these two attributes are often found together and are often considered necessary attributes of public goods, it is important for the present purposes to distinguish them.

The additional consumption of some goods, such as the further use of an idea, the additional viewing of a television program, the additional crossing of an uncongested bridge, or an additional observer of a high-wire act in an uncrowded setting, need not reduce the consumption of others. In these cases of classic nonrivalness, additional consumption has no marginal social cost, so it is a necessary condition of Pareto efficiency that the good not have a positive price. Such a good cannot be provided at the marginal social cost price of zero by a private firm that must cover its costs of production and cannot engage in price discrimination. This is true even when, as is often the case, nonpurchasers can be excluded at little or no cost. Even if some amount of the nonrival public good may sometimes be provided by a private firm that charges a monopoly price (or an average cost price), the less-than-optimal output that results from a nonzero price entails that the first-best conditions for Pareto efficiency are not met. In these cases, society faces a tradeoff between the losses from monopoly pricing and the losses from provision by imperfect governments with incentive-distorting taxation. Television is a conspicuous example of a nonrival public good subject (with scramblers) to exclusion that is sometimes privately provided at nonoptimal prices and sometimes provided by imperfect public agencies.

†*Discussants*: Rudolph Penner, U.S. Congressional Budget Office; Richard Musgrave, University of California-Santa Cruz; Gordon Tullock, George Mason University.

*Department of Economics, University of Maryland, College Park, MD 20742. I am grateful to Martin McGuire and Wallace Oates, and to the discussants of this session for most helpful criticisms, but I am solely responsible for any shortcomings.

Perhaps because nonrivalness and nonexclusion are usually assumed to go together, many economists have not noticed that nonrivalness is simply an extreme form of market failure due to economies of scale or decreasing costs. In *any* type of production in which the marginal cost is less than average cost at the marginal cost-equals-price quantity of production, there is market failure for *exactly* the same reason there is market failure for pure nonrival public goods for which exclusion is possible. Whenever the demand curve for any good cuts the marginal cost curve below the average cost curve, a firm charging the marginal cost price cannot recover its average costs without price discrimination, and thus is not viable.

At least when large numbers are at issue, markets also fail when it is uneconomic to exclude those who have not purchased a good from consumption of it. The concept of nonexclusion can also usefully be construed quite broadly. Any socially desirable redistribution of income or "social insurance" for which government is needed can be attributed to costs of exclusion. Adverse risk selection can in some cases make universal or compulsory social insurance appropriate, and this, in combination with the finding that the Friedman-Savage result on gambling is wrong and that the marginal utility of income characteristically declines with income (Martin J. Bailey et al., 1980), shows that redistributions of income to the poor that do not distort incentives too much will increase aggregate expected utility (my 1983 article). In any situation where a system of insurance has to be universal, exclusion is by definition ruled out. Just as any benefits of Richard Musgrave's (1959) "distribution branch" are a nonexclusive public good, so are any benefits of his "stabilization branch" of the government; everyone in the relevant economy tends to get the benefits of price stability and general prosperity whether he helped pay for them or not (Wallace Oates, 1972).

I. Exogenous vs. Endogenous Domains and Clienteles

The different public goods that correct different kinds of market failures have different patterns of beneficiaries. Most public goods have beneficiaries that are necessarily all within some geographical area or "domain." This domain may be either "exogenous" or "endogenous." If the domain in which the good is received is beyond the control of the political and legal system, I define it to be exogenous, but if it is determined by that system and its jurisdictional boundaries, it is endogenous to that system. Some public goods have a "clientele" that is not defined geographically. As we shall see, all public goods with clienteles rather than domains must have beneficiaries that are endogenous to the political or jurisdictional system. It will be convenient to begin with the simplest concept of the public good of exogenous domain and to assume initially that citizens do not move from residences in one jurisdiction to residences in another, but may commute anywhere within a metropolitan area.

The public good of air pollution control normally has an exogenous domain given by the relevant airshed; that of preventing a "greenhouse effect" through the burning of fossil fuels has an exogenous and worldwide domain because of the stratospheric winds. The boundaries of some other public goods and externalities of exogenous domain are given by ecological systems, by watersheds, by beautiful vistas, by the areas where accidental fires can spread from property to property, and by contagion zones (all the countries of the European continent are one zone at least for agricultural pests and diseases, but Australia is a separate zone). The boundaries of other public goods of exogenous domain are given by metropolitan boundaries, commuting distances, and social patterns that are not (except over a very long term) normally within the control of government. The boundary of the exogenous public good arising from the apprehension and incarceration of a criminal is determined by the locations of the crimes the criminal would have committed had he not been caught. Crime like many other social pathologies is mainly determined by commuting areas, so the metropolitan area is often the exogenous domain for many public goods and externalities. All public goods of exogenous domain

are *necessarily* goods for which exclusion of nonpurchasers is infeasible.

The nonexclusive public good of endogenous domain is admittedly a less-straightforward concept. It may seem that it is logically impossible to have a good for which exclusion of nonpurchasers is impossible, yet such that the government may set the boundaries within which it is received. If there were no limitations or constraints whatever limiting a jurisdiction in determining who received its services, this category would be empty. But there are constraints that give the government a range of choice in determining the geographical domain within which the public good is available, yet entail that it is nonexcludable in that area. Consider the relatively simple case of the public good of protection of property rights in real estate. A jurisdiction can protect such rights within its boundaries, but not outside them. It might seem that individuals could easily be excluded from the use of the courts and police that enforce property rights. But note that if my property rights in land in the jurisdiction are protected, and those of others in the same jurisdiction are not because the system is exclusive and not impartial, then it will not be rational for others to buy my property and it will then be worth less to me. Thus the costs of excluding individuals from the system of law and order within a jurisdiction are considerable even though the boundaries of the jurisdiction can be determined endogenously. More severe punishments by a particular jurisdiction for crimes within its legal boundaries similarly protect people nonexclusively within that jurisdiction, but may even increase crime in more lenient neighboring jurisdictions.

Though nonrivalness is often evident in nonexclusive goods, it also occurs separately. When it does, the possibility of exclusion necessarily makes the clientele endogenous. When these goods are at issue there is no need for all consumers in a given geographical area to have the same supplier, or that a given supplier restrict its supply to a given geographical area. Once produced, a television program may be viewed by any number of people in diverse areas without diminishing the value to existing consumers, but scrambling technology makes the exclusion of nonpurchasers possible. With nonrival public goods of endogenous clienteles, it is not surprising that both public and private provision occur.

New ideas that may be used by any number of people without diminishing their value to other users are nonrival public goods, even if they can be patented or copyrighted and nonpurchasers thereby excluded. New ideas of this sort characteristically arise because of both public expenditure and private entrepreneurship. When they can be effectively patented or copyrighted, they have endogenous clienteles; when they cannot, they have an exogenous worldwide domain.

There are also distinctive public goods that require separate discussion because they are relevant only in certain "catchment areas" that are determined by the pattern of preferences of the residents of the area. The patterns of preferences for public goods are dramatically dependent upon the language, religion, race, or (more generally) the culture of the people in an area. Most people obviously prefer political leaders and public servants who speak their own language, as is obvious from the histories of Belgium, Canada, Switzerland, and the successor states of the Austro-Hungarian empire. They similarly prefer governments of their own religion, ethnicity, or race, as is obvious from northern Ireland, the Middle East, and some large American cities. Over a sufficiently large number of generations, a government may create a culture that suits its convenience (France has become a country with a common language only since the end of the Middle Ages). But, in the very long run, even our grandchildren are all dead and so for practical purposes the culture of different groups must be taken as given; the taste of the Poles, for example, for distinctive and separate public goods was not erased even over several generations. When a group of people has a distinctive culture, much of the interaction, whether transfers of property rights or social pathologies, tends to be within the social group. This, along with distinctive tastes and the obvious preference for dealing with regulations in one's own language and officials of one's own culture, implies that

many of the distinctive public goods that peoples of varying cultural background want are public goods of exogenous domain.

II. Fiscal Equivalence

Whenever public goods have an exogenous domain, there are enormous advantages to government boundaries that match the exogenous boundaries of the domain of the public good. It might seem that considerations of economies and diseconomies of scale should also ideally influence the size of jurisdictions. But this is not so. The scale of the area of provision is not a variable of choice for the public good of exogenous domain. It could be that flood control has lower or higher average costs, per unit of population and property protected, with big rivers and flood plains than with small ones. But the size of the valley is given.

If the jurisdiction that provides a public good does not include all of the area in the exogenous domain of the good, there will be a spillover or external economy from the jurisdiction to neighboring jurisdictions. When the city-center government deals with air pollution or apprehends career criminals, it confers an external economy on the suburbs where some of the air pollution and criminal activity would have gone. Similarly, when Britain or the United States generate acid rain that also falls on Canada or Germany, there is again an externality that leads to less than optimal provision.

A sufficiently large government, such as a world government, would have no nonoptimality due to externalities. Many externalities are similarly avoided in large countries with unitary governments, such as France. But when the government is far larger than many of the public goods of exogenous domain, there is the political problem of the "internality" that also leads to nonoptimality. The gains from providing a local public good of exogenous domain can greatly exceed the costs of providing it, but, with a unitary national jurisdiction, the number of losers from the national taxes that would finance the public good will be far larger than the number of gainers. Thus the provision of the local public good will fail to command a majority of the larger jurisdiction. Representative political systems tend to provide some local public goods because of logrolling and other types of bargaining, but the difficulties of such bargaining are considerable and these processes often work badly. Thus there is also a grave disadvantage in having only a large unitary jurisdiction that is far larger than many of the public goods of exogenous domain.

There is accordingly much to be said for "fiscal equivalence" (see my 1969 paper), or jurisdictional boundaries that match the catchment areas of those public goods with exogenous domains. There is a need for jurisdictions that match pollution problems and other natural boundaries, and, in particular, a need for metropolis-wide jurisdictions. There is also a need for distinctive jurisdictions whenever peoples in different communities demand greatly different public goods and political leaderships. A lack of separate jurisdictions introduces a gratuitous uniformity in consumption when there are different preferences for public goods that reduces welfare (Martin McGuire, 1974). There will at the same time usually be other public goods of exogenous domain that transcend cultural boundaries, so fiscal equivalence normally calls for larger, pluralistic jurisdictions as well as smaller ones matching cultural communities. Switzerland is a country where arrangements of this kind have worked well. Canada probably would not exist at all were it not a federation with a separate province largely for Francophones.

Though fiscal equivalence calls for a multilevel mosaic of jurisdictions, it does not require separate bureaucracies or elected officials for every jurisdiction. There are economies of scope in the public as in the private sector. Often a public good of exogenous domain could be voted on (and, if chosen, paid for) only by the people in the domain, but the necessary administrative work could be handled by a jurisdiction or organization that achieved economies of scope. Some local governments apparently already provide for separate votes, tax assessments, and additional public services such as sewage or roads for distinct neighborhoods or development tracts.

In the very different case where there is endogeneity and also a clientele rather than a domain, the scale of any organization for provision should be given by the intersection of the demand curve and the marginal cost curve. With no problem of exclusion, it is an open question in each case whether there should be public or private provision. The losses from the absence of marginal cost pricing when nonrivalrous public goods are provided privately have to be weighed against the incentive-distorting effects of taxation and the imperfections of government: the market failure and the government failure need to be compared. An ideological claim that either public provision or private provision is right for all cases is unlikely to be correct. For some nonrivalrous goods, such as television, it is interesting that many countries use both public and private provision, and there is no theoretical reason why this is necessarily wrong.

With public goods from which nonpurchasers cannot be excluded but which nonetheless have an endogenous domain, the normative theory begins by suggesting that the economies and diseconomies of scale should determine the size of the jurisdiction. With governments as with private firms, there are indivisibilities in opportunity sets that give rise to economies of scale. There are also diseconomies of scale that arise because there can be only one source of coordination, and diseconomies of scale can occur when the supply of this coordination or management is combined with too large a supply of resources to be supervised effectively (Oliver Williamson, 1985). In space-intensive activities like agriculture and government, the costs of coordination are particularly great.

Consider now only local governments within metropolitan areas so that we may not only abandon the assumption that individuals do not change their jurisdiction of residence, but also suppose that such movement is costless. Let us also restrict ourselves to public goods with endogenous domains or with clienteles, so that the recipients of the public goods at issue can be determined by the political system and the sizes of jurisdictions determined by the economies and diseconomies of scale of the public good. In this set of circumstances, the famous Tiebout model applies. As Wallace Oates has helpfully pointed out to me, the movement of residence assumed in the Tiebout model effectively gives each jurisdiction an endogenous clientele, since the competitive jurisdiction can, by altering the character and level of its public good provision and the associated taxes, essentially choose its clients. As Rudolph Penner noted in his comment on this paper, the Tiebout model assumes that those with similar preferences for public goods move to the same jurisdiction, whereas this paper points out the advantages of drawing jurisdictional boundaries around pre-existing groups of similar preferences. These two approaches are complementary. The former can work well when mobility is inexpensive, as it may be within metropolitan areas over long periods, and the latter has the advantage when mobility is costly, as it is over larger areas and shorter time periods. The Tiebout model is not, however, applicable even at a local level to public goods that have a domain that is exogenous for reasons of nature or technology.

III. Military Economies of Scale

There is no reason to expect that the lowest point on the average cost curve will be the same scale for all functions of government. This consideration, like fiscal equivalence, argues against unitary governments and in favor of more differentiated governmental systems with a matrix of jurisdictions. The dramatic differences in the optimal scale of government for different functions can be illustrated by contrasting a good like protection against fires, which empirical studies have suggested can be provided efficiently on a small scale, with defense and military power. In general, an increase in the population of a country does not make it less secure. The costs of militarily defending the United States have surely not been increased because it has experienced population growth. Normally, a larger geographical area also does not add to costs of defense: Russia has been spared defeat by Napoleon and Hitler in part because it had a lot of space that permitted defense in depth and stretched

costs of defense and military power do not rise much, and may even diminish, with the population and size of a country.

Consider the per capita costs of a military capability costing, say, $200 billion. For a country of 200,000 population, the per capita cost of this military power would be an obviously unattainable $1 million per capita; for a country of 20 million, it would be $10,000, and for a country of 200 million, it would be $1,000 per head. Thus the economies of national scale in military power are staggering.

No wonder history is in large part a long story of aggression by big countries against smaller ones. The gains from exploiting the economies of governmental scale in military power are so striking that any number of kings and emperors have recognized them and used them to expand their domains.

These economies are so colossal that one must ask how there can possibly have been any equilibrium short of world government, whether arrived at by aggression, or by a peaceful Coasian agreement that would share the enormous savings among all the parties? The Napoleons of history have come close to world government; the Roman Empire, for a time, included most of the world it knew. The British Empire in the nineteenth century, though acquired in large part by classical-liberal governments interested above all in limiting government spending, came to include about one-fourth of the world's land area. So the economist must ask, why are there over a hundred independent countries in the world today and many tens of thousands of local, state, and special-purpose jurisdictions?

Part of the reason is surely the diseconomies of scale involved in coordinating and controlling vast spaces and numbers of people. But surely the advantages of something resembling fiscal equivalence have also played a role. The demand for separate jurisdictions for different cultural and linguistic groups that want distinctive and separate public goods, for separate jurisdictions that take account of other public goods of exogenous domain, and for jurisdictions that offer relief from the considerable diseconomies of scale in some local public services has been strong enough to overwhelm even the gigantic economies of national scale in military might. This suggests that there is some explanatory or predictive power in ideas of the kind that have been discussed here in a normative spirit.

This approach also suggests that the overwhelmingly large role of national governments, as opposed to both subnational and supranational jurisdictions, probably did not arise because of economies of scope or any other efficiencies. It has probably arisen mainly because national governments are the jurisdictions that have had the military or final power. This has given them the capacity to claim for themselves functions that often could have been performed more efficiently by other jurisdictions—often special-purpose jurisdictions—of both subnational and international character. Unitary national governments, in this view, are inferior to federalisms, and even federal countries could gain more from more decentralization and also stronger institutions for international cooperation.

REFERENCES

Bailey, Martin J., Olson, Mancur and Wonnacott, Paul, "The Marginal Utility of Income Does Not Increase: Borrowing, Lending, and Friedman-Savage Gambles," *American Economic Review*, June 1980, *70*, 372–79.

McGuire, Martin, "Group Segregation and Optimal Jurisdictions," *Journal of Political Economy*, January-February 1974, *82*, 112–32.

Musgrave, Richard, *The Theory of Public Finance*, New York: McGraw Hill, 1959.

Oates, Wallace, *Fiscal Federalism*, New York: Harcourt Brace, 1972.

Olson, Mancur, "The Principle of Fiscal Equivalence," *American Economic Review Proceedings*, May 1969, *59*, 479–87.

______, "A Less Ideological Way of Deciding How Much Should Be Given to the Poor," *Daedalus*, Fall 1983, 217–36.

______, "Space, Agriculture, and Organization," *American Journal of Agricultural Economics*, forthcoming.

Williamson, Oliver, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.

The Interaction of State and Federal Tax Systems: The Impact of State and Local Tax Deductibility

By DANIEL R. FEENBERG AND HARVEY S. ROSEN*

President Reagan's tax reform proposal calls for the elimination of the deductibility of state and local taxes. Critics of the president have argued that elimination of deductibility would put an unfair burden on residents of high tax states, harm the middle class, and have a disastrous impact on state and local public finance.[1]

Who would be most hurt if this part of the president's proposal were adopted? The purpose of this paper is to provide estimates of the impact of removing the deductibility of state and local taxes. We show how deductibility affects marginal and average tax rates for both state and federal tax systems. We provide relatively detailed information on the impact upon state income and general sales tax structures. Due to lack of appropriate data, we cannot consider property taxes in comparable detail, although we do take these taxes into account in the computation of federal tax liability.

Obviously, the potential impact of removing state and local tax deductibility depends upon what the rest of the tax code looks like. At this point, no one knows exactly what will emerge from the legislative process. We examine the impact of deductibility both under the status quo and under the president's proposal. These results should be of some use in assessing the implications of any "in-between" proposals that are presented.

Section I describes our data and methods. Section II shows how marginal and average tax rates for state and federal tax systems are affected by the deductibility of state taxes. One striking result is that combined federal income tax and state tax burdens would generally fall under the president's proposal, even for high-income individuals in high-tax states. A concluding section offers some brief comments on the political debate surrounding the deductibility of state taxes.

I. Data and Methods

The basic data source for this study is a stratified random sample of 38,000 federal income tax returns for the year 1982. (The computer file with these data is documented by Michael Strudler, undated). Most returns include the taxpayer's state. However, tax returns with Adjusted Gross Income (AGI) over $200,000 do not include a state identifier and are therefore excluded.

We have programmed the major individual income and general sales tax rules (which together comprise about 60 percent of states' revenues from their own sources) for every state for the year 1982.[2] With this information, we can estimate each taxpaying unit's state individual income and general sales tax liabilities. For purposes of simplicity, instead of reporting results for the income and general sales tax separately, we view them as two components of a single structure. Thus, for example, "the" marginal tax rate is the increment to the sum of income and sales taxes associated with a dollar increase in income. Unless otherwise noted, then, when we refer

*National Bureau of Economic Research, and Princeton University, Princeton, NJ 08544, respectively. This research is part of the National Bureau of Economic Research's Project on State and Local Government Finance. It summarizes results from our 1985 NBER Working Paper. We are grateful to the National Science Foundation, grant no. SES-8419238, for financial support, and to Larry Lindsey for allowing us to use the tax calculator for simulating the president's tax reform proposal.

[1] One critic, Governor Cuomo of New York, opined that eliminating deductibility would "pulverize... the middle class" (*New York Times*, November 24, 1985, p. 29).

[2] Details on the procedure are provided in our 1986 paper.

to "state tax structure," we mean the combined individual income-sales tax structure.

Our tax simulation model allows us to compute any desired summary measures of each state's tax structure under alternative tax regimes. Our focus is on marginal and average rates (with respect to *AGI*) faced by members of different income groups. To be more concrete, we adopt the following notation: T_s, T_f = state and federal tax liabilities, respectively; and t_s, t_f = state and federal *gross marginal tax rates*, respectively.

These are obtained by finding the incremental tax liability associated with a $1 increase in taxable wage income, and *not* taking into account the fact that state taxes can (sometimes) be deducted on federal income tax returns, and federal taxes can (sometimes) be deducted on state income tax returns. $I_f = 1$ if the taxpayer itemizes on the federal income tax return, and takes the value zero otherwise. $I_s = 1$ if the taxpayer can deduct federal taxes on the state return, and takes the value zero otherwise; and Y = Adjusted Gross Income.

In general, an individual's state and federal tax liabilities are nonlinear functions of income. Hence, we can write

$$(1) \qquad T_f = f(Y - I_f T_s - a_f),$$

$$(2) \qquad T_s = g(Y - I_s T_f - a_s),$$

whereas a_f represents reductions in taxable income (other than state income and sales taxes, but including local taxes) that are allowed in the computation of federal income taxes, and a_s is defined analogously.

In reality, deductions of tax payments are always done on a cash rather than liability basis. This avoids burdening the taxpayer with solving a system of nonlinear equations, but requires knowing the cash payments, which are not available to us. Therefore we approximate the cash payment with the calculated liability. In a steady state these should be identical, but the difference might be significant during the transition to a broad-based tax.

Consider now a $1 decrease (in absolute value) of a_s, that is, a change in the state tax law that increases state taxable income by one dollar. We define the state *net marginal tax rate*, τ_s, as the sum of the associated changes in state and federal tax liability. That is, it is the total increase in tax liability, taking into account the fact that changes in the state tax law have an impact upon federal tax liability.

In terms of the notation developed above,

$$(3) \quad \tau_s = -\left(\frac{dT_s}{da_s} + \frac{dT_f}{da_s}\right) = \frac{g'(1 - f'I_f)}{1 - I_s I_f g_i' f'},$$

where g_i' is the marginal tax rate of the individual income tax component of the state income-sales tax system. The presence of g_i' is due to the fact that when $I_s = 1$, only state income tax liability is affected.

The interpretation of equation (3) is straightforward. For individuals who do not itemize deductions on their federal returns ($I_f = 0$), the state marginal tax rate is determined entirely by the slope of the state tax structure, g'. For individuals who itemize on their federal returns ($I_f = 1$), the incremental tax burden is reduced by the federal marginal tax rate, f', times the increase in state taxes, g'. Hence the presence of $(1 - f'I_f)$ in the numerator. However, for individuals who also can deduct federal taxes on their state tax returns, the fact that federal tax liability has gone down creates a second-order increase in state tax liability.[3] This accounts for the presence of the term $(1 - I_s I_f g_i' f')$ in the denominator.

The federal net marginal tax rate, τ_f, is defined symmetrically,

$$(4) \qquad \tau_f = \frac{f'(1 - g_i' I_s)}{1 - I_s I_f g' f'}.$$

Average tax rates under various tax regimes are also of interest; these are defined in the obvious way as T_s/Y and T_f/Y for state and federal tax structures, respectively.

Before proceeding to the results, several limitations to our methodology should be

[3]Again, it is the steady-state liability that increases. In a literal sense, there is no change in liability for the current year, because of the cash basis for deductions.

noted:

(a) We do not allow for any behavioral response to tax code changes. Presumably, if deductibility were removed, states and localities would modify their spending and taxing decisions. (See Robert Inman, 1985, and Nonna Noto and Dennis Zimmerman, 1984.) Our results are therefore best viewed as estimates of the initial impact. The reason for neglecting behavioral responses is not that we think they are unimportant, but rather that estimating them in a reliable way would carry us too far afield. In this context, it is important to note that once deductibility is eliminated, any induced changes in state and local tax systems will have no feedback effects on federal revenues.

(b) Closely related to point (a) is the fact that our results tell us only about the statutory incidence of the various tax systems. Standard theoretical considerations suggest that economic incidence may be quite different. Having made this observation, we hasten to add that any serious study of the economic incidence of state and local tax deductibility must begin with careful analysis of its statutory impact.

(c) Our income variable is annual Adjusted Gross Income. For many problems, some indicator of permanent income is more appropriate.

(d) Our simulations are not revenue neutral. That is, when revenues are gained due to the removal of deductibility, we do not lower taxes elsewhere in the system in order to keep revenues constant. In the current political environment, it is impossible to predict with any confidence whether Congress would lower marginal tax rates, increase exemptions, or what. Indeed, adjustments might take place entirely outside of the income tax system in the form of changes in business taxes, or perhaps reductions in the deficit. In the face of such uncertainty, it seemed that our results would be most compelling if we simply refrained from hazarding a guess.

II. Results

As noted above, the potential impact of removing state and local tax deductibility depends upon what the rest of the tax code looks like. We begin by examining the impact of deductibility under the status quo, that is, the tax law as it existed in 1982. We then go on to study the impact of deductibility under the president's proposal.[4] (The president's proposal, which incorporates many modifications to the existing tax code, is described in detail in *The President's Tax Proposals to the Congress for Fairness, Growth, and Simplicity*, 1985.)

We first calculated the outcomes on a state-by-state basis, and then took means over states, weighting by the number of tax returns in each state. The results for all income classes are summarized in Table 1. Lines 1–8 refer to the status quo, that is, the federal and state tax laws as they stood in 1982. Line 1 shows t_s, the gross state marginal tax rate (*MTR*); line 2 shows t_f, the gross federal *MTR*. Net state *MTR*'s (τ_s) and net federal *MTR*'s (τ_f) are in lines 3 and 4, respectively. Average tax rates (*ATR*'s) for the state and federal systems are in lines 5 and 6, respectively; the corresponding *ATR*'s when state and local tax deductibility is disallowed are in lines 7 and 8.

Lines 9–11 pertain to the president's proposal. Specifically, we analyze the president's proposal as it would have applied to the year 1982.[5] Line 9 shows the gross federal income *MTR*. Line 10 shows the net federal income *MTR*, and line 11 shows the federal *ATR*.

The following are the main results that emerge from Table 1:

(a) From lines 1 and 3, we see that on average, under the status quo there is only one-half of a percentage point difference between gross and net state marginal tax rates. From lines 5 and 7, the mean difference between state average tax rates with and without deductibility is about 0.1 percentage points. Note that in effect, the figure in line 7

[4]Another interesting comparison would be between the president's plan and what that plan would be if amended to retain deductibility. This is done in our 1985 paper.

[5]All money values are deflated by the change in the *Consumer Price Index* for urban wage earners between March 1, 1981 and March 1, 1985. Because the president's proposal is completely indexed, this is an attractive method of using our 1982 data set.

TABLE 1—ALL INCOME GROUPS

Status Quo	
Gross MTR	
1) State	4.5
2) Federal	19.3
Net MTR	
3) State	4.0
4) Federal	19.2
ATR	
5) State	4.8
6) Federal	14.1
ATR without Deduction	
7) State	4.9
8) Federal	15.4
President's Proposal	
9) Gross Fed *MTR*	14.7
10) Net Fed *MTR*	14.6
11) Fed *ATR*	11.5

TABLE 2—RESULTS BY INCOME GROUP

	\$0–\$10,000	\$10,000–\$20,000	\$20,000–\$40,000	Above \$40,000
Status Quo				
Gross MTR				
1) State	3.0	4.6	5.8	6.6
2) Federal	8.5	19.7	27.0	38.2
Net MTR				
3) State	3.0	4.4	4.8	4.4
4) Federal	8.4	19.6	26.7	37.9
ATR				
5) State	3.7	4.3	4.8	5.2
6) Federal	3.1	9.6	13.6	20.5
ATR without Deduction				
7) State	3.8	4.4	4.9	5.4
8) Federal	3.2	9.9	14.7	23.2
President's Proposal				
9) Gross Fed *MTR*	5.3	16.4	20.0	27.7
10) Net Fed *MTR*	5.3	16.2	19.7	27.4
11) Fed *ATR*	1.4	7.4	10.9	16.8

would be the (gross and net) state average tax rate under the president's proposal.

(b) From lines 2 and 4, the ability to deduct federal taxes on state returns does not have much of an impact on federal marginal tax rates—the average difference between gross and net is only 0.1 percentage points.

(c) From lines 6 and 8, the mean decrease in federal average tax rates due to the deductibility of state personal income and general sales taxes and local taxes is 1.3 percentage points.

(d) From lines 2 and 9, under the president's proposal gross federal marginal tax rates fall substantially; so do net federal marginal tax rates (from lines 4 and 10). From lines 6 and 11, federal average tax rates also fall. None of this is too surprising given that lowering individual income tax rates is the centerpiece of the president's proposal.

(e) Due to space limitations, we cannot report results on a state-by-state basis. As we show in our 1985 paper, however, the small changes in means noted in items (a) through (c) above mask considerable differences across states. For example, while the mean difference between gross and net state marginal tax rates is 0.5 percentage points, for Minnesota and New York the corresponding figures are 1.4 and 1.9 percentage points, respectively. Similarly, while under the status quo the mean decrease in federal average tax rates due to deductibility is 1.3 percentage points, for Minnesota it is 2.1 percentage points, and for New York it is 2.6 percentage points. Having pointed out that differences do exist between states, we are left with the question of whether they are large enough to account for the enormous interstate differences in states' political reactions to the threat of removal of deductibility. We return to this question later.

A. *Results by Income Class*

We next consider the differential impact of deductibility by income class. The columns in Table 2 shows results for households with *AGI* in the \$0–\$10,000, \$10,000–\$20,000, \$20,000–\$40,000, and above \$40,000 ranges.[6] Examination of the table yields the following observations:

(a) Comparing the line 1 results for the various income groups, we see that on average, state tax systems are characterized by increasing gross marginal tax rates, and thus can be considered "progressive." But when state *net* marginal tax rates (from line 3) are considered, the following striking result occurs: not only do state net marginal tax rates increase more slowly with income than do their gross counterparts, but the net marginal tax rate for the over \$40,000 income class is less than that for the \$20,000–\$40,000 income class. Thus, the increasing incidence of

[6] We do not display separately summaries of the few returns with negative Adjusted Gross Income, although these are included in Table 1.

itemization at high-income levels plus increasing federal marginal tax rates "overcomes" the higher state statutory rates. However, it may be inappropriate to think about progressivity only for one component of the tax system as a whole; lines 3 and 4 make it clear that the state and federal marginal tax rates combined increase throughout the income scale.

(b) For the low-income groups, the issue of deductibility is not very important, simply because the proportion of itemized federal tax returns is so low. For *AGI* under $20,000, state gross and net tax rates, both marginal and average, are very close to each other, and the same is true for federal liabilities.

(c) In the upper-income groups, a high incidence of itemization together with high marginal tax rates leads to sharp divergence between net and gross marginal state tax rates under the status quo. For example, from lines 1 and 3 of Table 2, itemization for households with *AGI* in excess of $40,000 cuts state marginal tax rates by one-third, from 6.6 percent to 4.4 percent. For the same group, removing deductibility under the status quo would increase average federal tax rates by 2.7 percentage points, from 20.5 percent to 23.2 percent (see lines 6 and 8). Again, though, there are considerable differences across states which are not detailed here. New York is the most dramatic example. For households with *AGI* greater than $40,000, deductibility lowers state marginal tax rates from 18.8 to 11.3 percent, and lowers average federal tax rates from 25.0 to 20.1 percent.

(d) Under the president's proposal, for each income group the impact of removing deductibility of state and local taxes is overwhelmed by the reduction in federal tax rates.[7] Consider, for example, the figures for the very highest income group in Table 2. From lines 7 and 11, we see that for those with *AGI* in excess of $40,000, under the president's proposal the average tax rate for state and federal income taxes taken together is 22.2 percent. Under the status quo, the overall average tax rate is higher, 25.7 percent. (Add together lines 5 and 6.) As a proportion of income, then, the state and federal tax liabilities of high-income people fall under the president's proposal.

This result, of course, is for the average high-income household. What about those who reside in high-tax states? Calculations analogous to those presented in the preceding paragraph indicate that for high-income New Yorkers, overall average tax rates fall from 31.5 to 29.1 percent under the president's proposal; for Minnesota the president's proposal induces a slight rise, from 28.3 to 28.8 percent. Thus, even in states where one would expect high-income households to be very adversely affected, they either come out ahead like everybody else, or suffer very small increases in their overall average tax burdens.

III. Concluding Remarks

We have shown how effective state sales and income tax rates are affected by the federal income tax system and vice versa. The analysis has been conducted in the contexts of both the status quo and the president's proposal. In our discussion of the method for doing the calculations, we were careful to stress its limitations. At this juncture we would like to note again one of these limitations—the calculations are "static" in the sense that they do not take into account possible behavioral reactions by governments and households. The reason for emphasizing this point is because we think it helps explain a puzzle raised by our calculations. Namely, we have shown that the trade of state and local deductibility in return for lower federal tax rates (in conjunction with other provisions of the president's proposal) appears to be a good deal for all income groups in virtually all states. If that is true, why is the idea so controversial? Several possible explanations hinge on beliefs that people might have regarding behavioral responses to the proposal:

(a) People may not believe that federal marginal tax rates will stay at the levels in the president's proposal. A federal income tax base including state and local taxes may

[7]Of course, other aspects of the president's proposal, such as higher exemptions, also contribute to lower average tax rates.

ultimately lead to higher tax burdens than are possible under the status quo. In short, the lower average tax rates reported in our tables might be perceived as only temporary.

(b) The president's proposal includes higher business taxes. The people objecting to the plan may believe that the incidence of these taxes will be upon themselves.

(c) The figures in Table 2 imply that for high-income people, the tax price of state-provided goods and services would on average increase by 33 percent, and in a state like New York, it would be almost double that. No one knows exactly how such an increase in tax prices would change voter behavior. Perhaps some people who are against the proposal anticipate that it would generate substantial pressure to lower state and local government expenditures, and they view these expenditures as particularly desirable from a social point of view. Alternatively, opposition to the removal of deductibility may be coming from those who have a special interest in the existence of large state and local public sectors—elected officials, civil servants, public sector union leaders, et al.

REFERENCES

Feenberg, Daniel R. and Rosen, Harvey S., "The Deductibility of State and Local Taxes: Impact Effects by State and Income Class," NBER Working Paper No. 1768, October 1985.

______ **and** ______, "State Personal Income and Sales Taxes: 1977–1983," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press, 1986.

Inman, Robert P., "Does Deductibility Influence Local Taxation?," NBER Working Paper No. 1714, October 1985.

Noto, Nonna and Zimmerman, Dennis, "Limiting State-Local Tax Deductibility: Effects Among the States," *National Tax Journal*, December 1984, *37*, 539–50.

Strudler, Michael, "General Description Booklet for 1982 Individual Tax Model File," U.S. Treasury, Statistics of Income Division, Washington, undated.

President's Tax Proposals to the Congress for Fairness, Growth and Simplicity, Washington: USGPO, May 1985.

Budget Reform and the Theory of Fiscal Federalism

By JOHN M. QUIGLEY AND DANIEL L. RUBINFELD*

No matter how they are ultimately enacted, the federal budget reforms debated for the past two years will substantially alter the structure of fiscal federalism in the United States. Among the serious proposals are: 1) the consolidation and reduction of categorical and matching grant programs; 2) the elimination of the provision that allows state and local taxes to be deducted as personal expenses under the federal internal revenue code; and 3) the demise of general revenue sharing. To a large extent, of course, these specific proposals are motivated by the desire to reduce federal deficits or to reduce the size of the public sector of the economy. Not incidentally, however, the proposals are endorsed by a "revisionist" theory of fiscal federalism, the topic of this paper.

Much of the debate about tax and budget reform has been partisan, political and ideological in nature. The professional discussion of these components of the "New Federalism" among economists has been more muted, but not without contention. Of course, it is hardly surprising that professional economists disagree about the individual merits of proposed policy reforms, yet it is striking to note how the nature of the analytical arguments made and the rationales used to evaluate such reforms have changed over the past two decades. The state-local public finance of the 1960's was marked generally by an emphasis on the advantages of the federal government as a raiser of revenue, and as a corrective mechanism for market failures at the state and local levels. The state-local public finance of the 1980's is marked, in contrast, by a skeptism about the ability of the federal government to perform these functions, and also by a renewed emphasis on the presumed allocative efficiency arising from a system of multiple governments in which voting with one's feet is a serious option.

The current budget reform policy debate serves as a useful point for a selective review of analytical developments in local public economics. Recent developments do represent a shifting emphasis, but in our view, do not discredit the earlier consensus. A taxonomy of recent developments includes: 1) Tiebout models, which view the state-local public sector in the context of a static long-run equilibrium model devoid of politics, but in which individuals are mobile among numerous jurisdictions; 2) Leviathan public choice models, in which those making tax and budgetary choices (or setting agendas for voters) have substantial economic and political power over inputs, outputs, or agendas; 3) general equilibrium models of taxation applied to local government, especially the application of the Harberger model to study the incidence of property taxes. A final development, which represents an embryonic framework for study, is based on dynamic game-theoretic models in which state and local governments compete, but in which limited information and the presence of externalities make the existence of equilibrium problematic, and its efficiency questionable.

I. Categorical and Matching Grants

Historically, a substantial proportion of all federal grants-in-aid to state and local governments has consisted of categorical or matching grants, rather than general purpose or block grants. However, since 1980 there has been a concerted effort both to consolidate matching programs into block grants, and to reduce the magnitude of federal grants programs (Robert Inman, 1985). For example, categorical aid in 1983 was $28.8 billion, a marked decrease from its 1972 level (in 1983 dollars) of $44.2 billion. Block grants

*University of California, Berkeley, CA 94720. This essay was improved by the suggestions of George Break, Robert Inman, Richard Nathan, and Wallace Oates.

had increased, however, from $6.7 billion to $12.9 billion.

Prior to the 1960's, the public finance equivalent to the Coase Theorem was dominant. According to that view, most externalities involved interactions between neighboring jurisdictions and were well understood by those involved. Under such circumstances, it was plausible to conclude that any gains from trade due to externalities could be realized directly by negotiation. Direct federal intervention was unnecessary, since a process of bargaining under conditions of full information generates efficient outcomes.

The Coasian view gave away to an interventionist prescription when externalities involved were perceived to be operating on a larger scale across many jurisdictions, generating externalities that were not bilateral in nature. As a result, the strong federal budgetary reliance on categorical and matching grants during the 1960's and 1970's was supported strongly by the economics literature at the time. Models of federalism tended to view the economic choices of jurisdictions as if made by a single individual, in which the behavior of the mayor or the median voter is modeled analogously to the behavior of a consumer in the private economy (Wallace Oates, 1972). According to this model, the median voter responds to the price reductions of matching grants according to the standard analysis of income and substitution effects. Because there are spillovers among jurisdictions, however, this system of matching grants is capable of improving social welfare by removing the distortions due to externalities. Matching rates for local spending are chosen to equal the fraction of local output which spills out to other localities. Underlying this strong interventionist view are two assumptions: 1) there is full information about the level of the externalities as well as production and consumption opportunities; and 2) local government choice is governed by the price and income elasticities of citizen demand for public output.

By the 1980's the decentralized solution to the problem of interjurisdictional externalities had been brought under direct attack. In part, this attack arose from the recognition that levels of externalities are difficult to measure, and that it is even more difficult to legislate matching rates that resemble the rate of spillout. More generally, however, the model of a passive, optimizing local decision maker had become suspect. The broad set of matching grant programs had generated a new industry—grantsmanship—in which local bureaucracies expend substantial resources and negotiating effort to obtain categorical funds or favorable matching rates.

Of more theoretical importance, however, by the 1980's much of the public choice literature attacked the median voter or decisive decision-maker model on a broad scale. Leviathan—the self-aggrandizing politician or decision maker whose objective is to maximize budget size rather than citizen utility —was growing. At the local level, Leviathan's power is exerted in a number of different forms, from control over entry (James Buchanan, 1965), to direct control over budgets and the information presented to citizens (William Niskanen, 1971), to referenda agenda setting (Thomas Romer and Howard Rosenthal, 1979), and through the exertion of public employee market power over wage rates (Paul Courant et al., 1979a). Initial attempts to model Leviathan suffered from a failure to take into account the possible responses of citizens who might be inappropriately treated. However, more recent modeling of the monopoly models of government has shown that neither the possibility of repeated referenda, nor the prospect of citizen migration to other jurisdictions, can fully eliminate Leviathan's disproportionate power (Courant and Rubinfeld, 1981; Dennis Epple and Alan Zelenitz, 1981).

The theoretical evaluation of matching grant programs to local government under the "old" and the "new" public choice models of government behavior could not be more apposite. Under the former view, the central question is to determine the appropriate matching rates to increase allocative efficiency, while conceding that lack of information puts us in a second-best world. According to the latter view, attempts at such "fine tuning" (to use an ancient metaphor) are difficult at best, and more likely to be counterproductive, given the Leviathan behavior of recipient governments.

II. General Revenue Sharing

Notwithstanding the macroeconomic arguments for revenue sharing, "fiscal drag," the microeconomic case was grounded in the traditional public finance-taxation literature which favored federal, as opposed to local, tax instruments on grounds of vertical equity, allocative efficiency, and even X-efficiency. According to the prevailing view of the 1960's, the federal income tax is mildly progressive, while the local property tax is a regressive excise on housing (Dick Netzer, 1966). Moreover, heavy reliance on revenues from a single excise tax instead of a broadly based income tax generates larger deadweight losses. Finally, federal taxation was argued to be desirable on administrative grounds. In contrast to the relative efficiency of the Internal Revenue Service in collecting revenues, local property tax administration was argued to be costly, and to involve assessment practices which generated substantial horizontal inequities (Walter Heller et al., 1968).

Finally, general revenue sharing to localities was viewed as an instrument to promote horizontal equity, since its allocation could be determined on a formula basis to reduce disparities in tax-price levels or to reward tax effort (Richard Musgrave, 1961). The most visible horizontal inequity occurred in the financing of local schools, it was argued. Under a local property tax base, identical households in different jurisdictions may pay substantially different unit prices for local schooling, due to the variation in the average house values and nonresidential property across school districts. The same argument could be applied among states to reduce differentials in the cost of government or in wealth-determined spending levels.

The federalism literature of the past decade, in contrast, has provided an alternative set of economic models and arguments which support the elimination of the revenue-sharing program altogether. One important development adds a note of discord to the horizontal equity arguments for revenue sharing. The use of grants to equalize tax prices among jurisdictions only makes sense as a policy if there is something to be equalized. The recent literature on tax capitalization suggests that there may not, since jurisdictions with substantial nonresidential property are more desirable as residences, other things equal, than jurisdictions with low levels of nonresidential property. Mobility and the ensuing capitalization cause property values to rise in high-tax-base jurisdictions and to fall in those with low-tax bases. As a result, tax prices differ among jurisdictions, but so do the "entry fees," the premiums paid for residence in different jurisdictions. In equilibrium, the entry fees plus the capitalized value of the tax-price differentials will equalize among jurisdictions, so that the equity imperative for tax-equalizing revenue sharing is gone (Bruce Hamilton, 1976).

More striking is the alternative view of vertical equity. According to the general equilibrium analysis of the property tax, the system of local property taxes is best seen as a national levy on real capital (Peter Mieszkowski, 1972). Like any broad-based tax, on inelastically supplied capital, a real property tax will be progressive in its incidence. According to this view, the system of local property taxation may be quite desirable on equity grounds.

Alternatively, the modern "Tiebout view" of the property tax is as a local benefits tax. This benefits taxation view of the local public sector grew out of Charles Tiebout's suggestive analogy (1956), but was really not articulated until much later. According to Tiebout's long-run equilibrium model, mobility can under certain conditions generate an efficient outcome in the market for publicly provided goods. This, of course, vitiates general revenue sharing—if the property tax is a benefits tax, then its progressivity is irrelevant. To the extent that one cares about incidence and vertical equity, the pattern of consumption of local public goods, and not the national distribution of the ownership of capital, is relevant under the Tiebout view.

A third argument against revenue sharing again relies on Leviathan, but in a somewhat different form. According to the "flypaper" theory, as coined by Arthur Okun, money (from the federal government) sticks where it hits. For example, local bureaucrats are likely to control and to spend more revenue-sharing funds than they would spend out of an equivalent increase in local resources. Em-

pirical evidence supporting the flypaper argument (Edward Gramlich, 1972), and theoretical papers showing how "flypaper" might be consistent with a limited information equilibrium (Courant et al., 1979b; Oates, 1979), all suggest that revenue sharing could contribute to budget maximization rather than utility maximization. Once again, the public finance literature of the most recent decade appears to provide support for the proposed restructuring of our federal system.

III. Deductibility

State and local taxes have always been deductible from personal income for purposes of federal income taxation as have most excise taxes. The economic justification for deductibility arises directly from the role of federalism in treating ability-to-pay taxation. As long as state and local taxes are seen primarily as ability-to-pay taxes, then according to the classical Haig-Simons view of taxation, the appropriate federal tax base is individual income less the ability-to-pay taxes imposed by other governments (George Break, 1980). The reasoning is clear. Only discretionary income should be taxed. If state and local taxes are raised in a nondiscretionary manner, these taxes should not be included in the base available to the federal government. The absence of discretion arises in part because, according to the traditional view, individuals do not have the option to avoid taxes, and in part because the benefits obtained from public services bear little if any relationship to taxes paid.

Implicit in this argument for deductibility is the view that local publicly provided goods and services are largely public, in the sense that they are nonrivalrous, not congested, and to a substantial extent, not excludable. In contrast, a substantial body of empirical public finance literature concludes that an alternative perspective is more appropriate. According to this newer view, most locally provided goods are available at constant costs. Roughly speaking, public goods are private goods that are provided collectively because exclusion is difficult, and because it is cheaper administratively to manage the provision publicly rather than privately (T. E. Borcherding and R. T. Deacon, 1972; T. C. Bergstrom and R. P. Goodman, 1973). The ease of administration arises in substantial part when the primary source of taxation is the local property tax, and the benefits of local public services are roughly proportional to house values.

To the extent that one is willing to view the local property tax as a benefits tax, the traditional argument for deductibility makes no sense. If state and local taxes are benefit taxes, and individuals are mobile among jurisdictions, then choices of public goods are just like choices among private goods. Therefore, state and local taxes are discretionary payments and should be subject to taxation at the federal level. If they are deductible, the tax system generates substantial inequity; the federal subsidy increases with income, and richer jurisdictions are likely to make larger tax efforts. Federal taxation is, in fact, necessary for an efficient, nondistorting set of local taxes to be levied—deductibility would generate inefficiences by lowering the effective tax-price of local public goods, thereby leading to overspending at the local level. Efficiency requires "many" communities and benefit taxation, which can be achieved if entry is restricted to require tax payments to equal the average cost of publicly provided goods (Hamilton, 1975).

As a number of authors have suggested, however, the efficient Tiebout equilibrium exists only under circumstances in which the publicly provided good is essentially a private good. The Tiebout model does not specify the source of taxation to finance the public good, but it might as well be a system of private user chargers, which make the link between payment and benefits generated explicit. Therefore, the Tiebout model supports the argument for the elimination of deductibility. The removal of deductibility will encourage local jurisdictions to switch to user charges or to the direct private provision of many services now provided collectively.

IV. Concluding Comments

We have suggested strongly that the widely discussed proposals for restructuring our federal system are supported by several strands of recent academic research. It seems

clear to us that this intellectual support could not have been given two decades ago. The current deductibility of nonfederal taxes is challenged by models in which local taxes are benefit taxes for publicly provided goods, and by general equilibrium models of local taxes, in which property taxes are more progressive than income taxes. Arguments for general revenue sharing based on tax price equity are similarly challenged by Tiebout models in which variations in tax rates are fully capitalized into property values.

Categorical and matching grant programs, as well as revenue sharing and deductibility, are challenged by Leviathan models in which budget maximers use price reductions or income increases to enlarge the public sector by more than is consistent with the price and income elasticities of citizen demands.

All three of these strands of theoretical analysis argue strongly for the kinds of fiscal reforms currently proposed. It would be seriously misleading, however, to ascribe to these newer theories a coherent, or even consistent, view of the state-local public economy. The conflicts and inconsistencies among these recent developments are striking. On tax incidence, for example, the "new view" of the property tax is inconsistent with the Tiebout view, despite the general equilibrium character of both theories.

The Tiebout and Leviathan models are thoroughly inconsistent with each other and have very different normative implications. Decentralization is desirable from the point of view of both models. In Tiebout models, decentralization allows government functions to be matched with the jurisdictions that are best able to perform those functions; in Leviathan models, decentralization provides an important constraint on the growth of government. Yet, the very mobility which leads to the Tiebout result that local taxes are benefit taxes and that tax variations are capitalized also implies that Leviathan bureaucrats will be unsuccessful. Conversely, the lack of ready alternatives which permits Leviathan government to extract resources from the citizenry means that tax prices will not, in general, be equalized by capitalization and that local taxes will not be benefit taxes.

You can't have it both ways.

REFERENCES

Bergstrom, T. C. and Goodman, R. P., "Private Demands for Public Goods," *American Economic Review*, June 1973, *63*, 280–96.

Borcherding, T. E. and Deacon, R. T., "The Demand for the Services of Non-Federal Governments," *American Economic Review*, December 1972, *62*, 891–901.

Break, George F., "Tax Principles in a Federal System," in Henry J. Aaron and Michael J. Boskin, eds., *The Economics of Taxation*, Washington: The Brookings Institute, 1980.

Buchanan, James M., "An Economic Theory of Clubs," *Economica*, February 1965, *32*, 1–14.

Coase, Ronald, "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

Courant, Paul N. and Rubinfeld, Daniel L., "On the Welfare Effects of Tax Limitation," *Journal of Public Economics*, December 1981, *16*, 289–316.

______, **Gramlich, Edward M. and Rubinfeld, Daniel L.**, (1979a) "Public Employee Market Power and the Level of Government Spending," *American Economic Review*, December 1979, *69*, 806–17.

______, ______, **and** ______, (1979b) "The Simulative Effects of Intergovernmental Grants: Or Why Money Sticks Where It Hits," in P. Mieszkowski and W. H. Oakland, eds., *Fiscal Federalism and Grants-in-Aid*, Washington: Urban Institute, 1979.

Epple, Dennis and Zelenitz, Alan, "The Implications of Competition Among Jurisdictions: Does Tiebout Need Politics?," *Journal of Political Economy*, December 1981, *89*, 1197–217.

Gramlich, Edward M., "Intergovernmental Grants: A Review of the Empirical Literature," in Wallace E. Oates, ed., *The Political Economy of Fiscal Federalism*, Lexington: D.C. Heath, 1972.

Hamilton, Bruce W., "Capitalization of Interjurisdictional Differences in Local Tax Prices," *American Economic Review*, December 1976, *66*, 743–53.

______, "Zoning and Property Taxation in a System of Local Governments," *Urban Studies*, June 1975, *12*, 205–11.

Heller, Walter W. et al, *Revenue Sharing and the City*, Baltimore: Johns Hopkins Press, 1968.

Inman, Robert P., "Fiscal Allocations in a Federalist Economy: Understanding the New Federalism," in John M. Quigley and Daniel L. Rubinfeld, eds., *American Domestic Priorities*, Berkeley: University of California Press, 1985.

Mieszkowski, Peter M., "The Property Tax: An Excise Tax or a Profits Tax?," *Journal of Public Economics*, April 1972, *1*, 73–96.

Musgrave, Richard, "Approaches to a Fiscal Theory of Political Federalism," in *Public Finances: Needs, Sources and Utilization*, Princeton: Princeton University Press, 1961.

Netzer, Dick, *The Economics of the Property Tax*, Washington: The Brookings Institute, 1966.

Niskanen, William, *Bureaucracy and Representative Government*, Chicago: Aldine-Atherton, 1971.

Oates, Wallace E., *Fiscal Federalism*, New York: Harcourt, Brace, Javanovich, 1972.

______, "Lump Sum Intergovernmental Grants Have Price Effects," in P. Mieszkowski and W. H. Oakland, eds., *Fiscal Federalism and Grants-in-Aid*, Washington: Urban Institute, 1979.

Romer, Thomas and Rosenthal, Howard, "Bureaucrats vs. Voters: On the Political Economy of Resource Allocation by Direct Democracy," *Quarterly Journal of Economics*, November 1979, *93*, 562–87.

Tiebout, Charles M., "A Pure Theory of Local Expenditures," *Journal of Political Economy*, October 1956, *64*, 416–24.

ROUNDTABLE ON ECONOMIC EDUCATION: INCREASING THE PUBLIC'S UNDERSTANDING OF ECONOMICS

The Marketplace of Economic Ideas

By ALBERT REES*

Since the title of these remarks is not very informative, I should begin by stating what I plan to discuss. My topic is the education of the adult public in economics through books, newspapers, magazines, and television.

There is still a solid core of agreement among American economists about a large part of economics—particularly about microeconomic theory and its applications. Much of this core of agreement has important policy implications, and differs from the preconceptions of the public and of elected officials about the way markets work. The oil price increases of the 1970's are a case in point. Most economists knew that the demand for oil is not completely inelastic, and foresaw that given time for adjustment, consumption would decline, exerting downward pressure on prices. The public, elected officials, and many bank lending officers seemed to believe that petroleum products are necessities, that therefore the demand for them is completely inelastic, and that prices that go up will stay up. Many of these people have learned to their sorrow that they were wrong, and some have lost fortunes in the process.

Economists also agree about the fundamental aspects of international trade theory and the principle of comparative advantage. They would almost all argue that a protective tariff on shoes will raise the price of shoes to American consumers, and would predict that if the tariff saves jobs in the domestic shoe industry, it will do so at an excessive cost. Neither the public nor the Congress shares this perception. On the contrary, they appear to believe that high tariffs are the only way in which one can protect the American working man from the unfair competition of low-paid foreign labor.

Economists do a good job of teaching the agreed parts of economic theory to college students, particularly those who take more than one course in economics, although students who pass their final exams may later forget what they have learned. However, economists do a very bad job of communicating the principles on which we agree to the general public through trade books, newspapers, and television.

The focus of public attention in economics is on macroeconomics, where policy decisions are centralized and highly visible. Every educated adult is concerned with the rate of inflation and the rate of unemployment, and many follow such esoteric measures as leading indicators and M_1. But in macroeconomics, the economics profession is now deeply and bitterly divided. The split between Keynesians and monetarists is, of course, of long standing, though Keynesianism was the dominant school until quite recently. Now, however, there are also supply siders, rational expectationists, and post-Keynesians, not to mention several minor schools of thought. Serious newspapers carry long accounts of the disputes among economists of these various views, and rival economists berate one another in the letters to the editor. The newspaper reader no more understands the differences between these schools than he understands those between the rival factions of the Palestine Liberation Organization. His impression is that economists are divided on all issues into irreconcilable warring sects. Since economists themselves cannot agree on anything, he reasons, no attention should be paid to them.

*President, Alfred P. Sloan Foundation, 630 Fifth Avenue, New York, NY 10111. I am indebted to my colleagues Arthur L. Singer, Jr., and Michael S. Teitelbaum for helpful comments on an earlier draft.

The situation is no better for other media. Television talk shows feature debates between representatives of opposing schools, often drawn from the extremes of the profession. A few television programs, such as Adam Smith's Money World, represent serious attempts at economic education of the public though even here areas of agreement may receive less attention than areas of dispute. Successful trade books in economics, unlike successful textbooks, are those that advocate some radical reform of the economic system, of which a return to the gold standard is one of the mildest, or those that predict the imminence of hyperinflation, the coming total collapse of the economy, or both.

Agreement among economists is simply not news. A statement released earlier this year advocating closing the budget deficit through a combination of reductions in domestic and defense expenditures and an increase in taxes was signed by seven former chairmen of the Council of Economic Advisers who had served three Republican and three Democratic presidents. It received scant attention in the press.

Another aspect of public attention to economics is that the media emphasize short-run forecasting. "Will we have a recession next year?," they ask economists, or "What will the Treasury bill rate be in March?" This emphasis leads to a view of economists as weathermen who tell you when you should carry your umbrella, and are blamed if it rains on what was supposed to be a sunny day. Unfortunately, economists are not very good at short-run forecasting—perhaps not even as good as meteorologists. When their forecasts differ, or the consensus forecast is wrong, the public concludes that economists have nothing to offer.

Among the economists who receive prominence in newspapers and television coverage, some are university professors or come from nonprofit research institutions, some are government officials, and some are employed by manufacturing corporations, banks, consulting firms, or trade unions. These economists share, for the most part, a common training, but they spend their working lives in very different ways and thus develop different interests and perspectives. Greater public awareness of these differences would help the public evaluate the views presented to it.

There have been several efforts to improve the media coverage of economics through programs to train economics journalists. These include the present Bagehot Fellows program at the Columbia University School of Journalism, and the former program at the Woodrow Wilson School at Princeton, which was funded by the Sloan Foundation. These programs are successful as judged by the reports of participants in them. It is very hard, however, to judge their ultimate effect on public understanding of economics. They reach only a small portion of the journalists covering economic affairs, and have a high cost per journalist trained. Publishers and broadcasters have not been willing to contribute much of this cost.

I should like to make very clear that it is not my intention to criticize the media. Interest in conflict rather than in harmony extends far beyond economics. Indeed, it is inherent in the definition of news, which is why we read more about Lebanon than about Switzerland. Rather I want to ask whether there is anything the economics profession can do about the problems it faces in communicating with the public.

One possibility is that we could settle the disputes over macroeconomics that now divide us. This seems to me to be unlikely in the foreseeable future. The natural sciences tend to have disagreements only at the frontiers of their disciplines. Behind these frontiers are large areas of settled theory that have been thoroughly tested. Disputes are settled by crucial experiments or crucial observations, usually replicated by several experimenters or observers, which leave little doubt as to whether a theory has fulfilled its predictions. Economics has only limited ability to experiment, and great difficulty in specifying and making crucial observations. Our disputes are therefore likely to be with us for many years.

Short of settling disputes over macroeconomics, I think we can still do a better job of communicating. Much of our problem arises because the profession properly spends its energy largely in talking to itself. Economists

Communicating Economic Ideas and Controversies

By LEONARD SILK*

I suppose I have thought about this subject more than any other in my professional life. It might shock you, therefore, when I say I don't think I have anything particularly useful or fresh to say about it. Occasionally, people tell me how much they've learned about economics from me. I'd like to believe them, and sometimes I do. On the whole, though, I don't notice any improvement in public understanding. Some people I meet say, "I read you all the time," while others say, "I read you, but I don't always understand you." Sometimes I respond, not insincerely, "I don't always understand myself." Of course, some people I meet say absolutely nothing. They're the ones who have decided to stay out of the Business Section altogether and enjoy life in the wonderful nonfinancial world.

So, despite a certain amount of complimentary mail and *viva voce* compliments, I am not all sure that I have raised the average level of the public's economic literacy one jot or tittle. The only thing that comforts me is that I don't think anybody else has either. But what if—as occasionally happens on good days when the writing has gone reasonably well or it has been well received—I feel that I do know something about what I'm doing? What if I thought you or somebody else were wrong? Is there anything I could tell you about it that you don't already know? Could I tell it in a way that does not sound like bragging?

We are all in this game together, trying to educate the public on economic ideas and controversies. None of us can afford to be smug or confident or proud about what we have managed to do. I do not know any statistical evidence, but I strongly suspect—judging from the public debates on economic issues, ranging from the farm problem to fiscal policy to free trade versus protectionism (after a couple of centuries of trying to get politicians and affected groups to think like orthodox economists on that last issue)—that the public's ability to deal with such controversial questions has not been significantly improved despite the torrents of ink and floods of words lavished on them by economists.

Why is that so, if it is so? I offer the following hypotheses:

1) Economists do not know how to communicate with the general public. They spend too much time in places like this, talking to other economists, or in classrooms, talking to young and innocent captive audiences who cannot escape with impunity in less than a semester or wreak vengeance on the instructor except by falling asleep.

2) The problem is the public's fault: The public is lazy, and it hates economics because it doesn't understand economics and because economics is too abstract and complex and interconnected, and the public won't wait around long enough to hear the whole story.

3) Economists aim, or think they aim, to serve "the public interest," whatever that is; but individual members and groups of the public are mainly concerned (as the same economists conventionally insist) with their self-interest. And the economists applaud that dedication to self-interest, as long as it manifests itself in the marketplace; they like it much less when it operates through the political process. But ordinary people find it harder to make the distinction between the market place and politics. Self-interest is self-interest, and it is the exceptional businessman or worker who thinks it's okay in one place and not in the other, especially when the businessman or worker is in serious trouble, in which case he has a good deal of difficulty understanding the economist who is urging him to drop dead or at least go bankrupt and gracefully grow old unemployed.

*Leonard Silk, Economics Columnist, *The New York Times*, New York, NY 10036.

4) Economics is a flawed science, and the public knows it. Economists generally accept the Friedmanian dictum that the test of a valid theory is its ability to produce correct predictions, but they routinely and very publicly fail that test, monetarists included. They say double-digit inflation is around the corner because the money supply has been growing much too fast, but when the double-digit inflation doesn't come but instead inflation slows down, they hardly ever recant and adopt new theories. Of course, there are exceptions, but by and large, if you took the positions held by economists at the moment they received their Ph.Ds and where they stand today, the redistribution of theoretical positions would not be remarkably large. The public can be forgiven for thinking that economic theories are held like religious dogmas, and few economists have the guts to become renegades or heretics from the faith they acquired in graduate school. And the public, without knowing the Friedman test, puts economists down for the frequency of their failed forecasts, and regards this as evidence of a failed science. Is this a bum rap? The normally wide variance of forecasts over any fairly long period ahead, such as a year or more, is evidence of the uncertainty of economic predictions and the theories on which they are based; the errors of the consensus forecasts, viewed after the fact, are further evidence of the discipline's weakness; and the fact that there is only one winner in the annual forecasting derby, but never the same winner, is additional confirmation of the hazard or luck involved in being a winning forecaster. By pretending to do what they cannot reliably do, economists have lost the respect of both the broad public and influential public and private decision makers.

5) Economists are as politically biased as most other people, but they hide their biases behind a scientific curtain. This is not necessarily done from a deliberate and conscious desire to deceive, since the most deep and dangerous biases of economists or anyone else are the unrecognized ones. But the public does recognize the biases of different economists and discounts their "scientific" or "positive" analyses for bias; as a result, the public downgrades economics as a pretender-science, especially as the economists disagree, each in the name of economic science.

6) My final hypothesis, at least for today, is that economists are stuck with an impossible task in trying to communicate with the general public, especially on controversial subjects. Economics is an extremely difficult and complex subject. Simple models may be necessary but, as applied to individual cases, they are frequently misleading or too skimpy to provide an adequate explanation. Complex models are too complicated for ordinary people; and raw empiricism tells you nothing of use for next time. So the economist may face the choice of oversimplifying and distorting, or of so weighing his explanation down with complexities and qualifications as to lose his audience by being incomprehensible and tedious. Some choice!

What can be done about all this? I don't know. If my impossibility hypothesis is correct, nothing. But, despite fits of despair, I am unwilling to accept that hypothesis. I go on writing for the public. Why? Standard economic analysis would say, "To make a living." Maybe standard analysis is right. Dr. Johnson said, "No man but a blockhead ever wrote except for money." Yet I persist in believing that my calling, our calling, is a noble one.

And I do feel very strongly that we must go on trying to fulfill our obligations to the public. So I offer you the following advice, grouping my suggestions as responses to my six hypotheses for the failures of economists to educate the public:

1) You cannot communicate with an audience if you cannot hold an audience. My colleague, Richard Shepard, a reviewer of books, movies, the Yiddish theater, and much else for *The New York Times*, recently wrote that the TV producer, Don Hewitt, "had the idea in the late 1960's of giving television viewers a program of short, lively pieces instead of the then prevalent one-subject shows that were overwhelming audiences with edification."[1]

[1]*The New York Times*, December 25, 1985, p. 29. The program Hewitt created was "60 Minutes."

Stop overwhelming people with edification!

A related message: Stephen Potter, author of *Gamesmanship*, observing the pompous and portentous way that economists said obvious and trivial things, called economics "the plonking science," which it still is.

Stop plonking!

2) What should we do if the public is too lazy or unable to understand economics and grasp the real issues in economic controversies? Pretend that it isn't! That is the only basis for a free and democratic society. If you can really believe that you or anybody else can't fool all the people all the time, so much the better. I believe that. If our aim is to educate the public, we must respect the public or wind up trying to trick and propagandize it. I learned in the Army a long time ago that most people are as smart as I am, and many are smarter, even if I know some things they don't know—yet.

3) I see the conflict between the economist's desire to serve the public interest and the tendency of individuals and interest groups to serve self-interest first as a real obstacle to economic education. I have no easy answer to it. However, I think most people are not just self-interested but care about others, often about the nation or even the human race. How can they not? No man is an island. Self-interest overlaps and is affected by the general interest, if not in the short run on a particular issue then in the long run on a wide range of issues. Almost by definition, most people will be worse off if the nation is weakened, most better off if the nation is strengthened. It's the economist's job to help the public understand the confluence of self, national, and international interests, and the moral dimensions of economic problems. I think economics should rid itself of its narrow and mistaken historical hedonism. As it is, I think many economists contradict and undermine the public interest they affect to serve. But the opposite danger also exists: that, in the name of the public interest, they sometimes ignore the damage to particular groups of people, and they do it with great confidence and aplomb when these are not the groups with whom they are allied by politics, interest, or sentiment.

4) Economists need to be more modest about their ability to predict the future. Just because the public, including business executives and highly placed politicians, wants them to make forecasts, and one-handed, unconditional forecasts at that (while contradictorily putting the economists down for their inability to forecast accurately), doesn't mean the economist has to play the role in which the public would cast him: first, as seer, then as phony.

But I do not say that economists should never forecast. Since much decision making in business and government is future-oriented, assumptions about the future must be made, and the economist who specializes in forecasting should help make them. The forecasting record of economic forecasters is imperfect and somewhat erratic, but it's a good deal better than nothing, and it gives a kind of base for decision and policy making. Can it be made much better? I'm dubious.

The problem is the inherent uncertainty of economic events. I don't see how economics, a wide-open science dependent on a host of variables whose relations are constantly changing, can ever close its system and standardize human behavior enough to become a reliable predictor of economic events and systemic conditions. Part of the economist's communication problem is to get the public to understand and accept that fact of life, and to accept what the economist can do as the best he can do, and better than random guesswork or idiot trend-setting.

5) The problem of political or ideological or interest-group bias on the part of the economist cannot be solved. I would not want it to be; economists are people, citizens, frocked or unfrocked philosophers, and they are entitled to their biases, that is, their values, like anybody else. But in my view, they have an obligation to know their own biases and to make them clear so the public can better understand and judge their analyses and policy recommendations.

6) I formally reject the proposition that increasing the economic understanding of the public is an impossible task. I just wish I had stronger empirical evidence on which to base that rejection. Nevertheless, I find the job of educating the public on economic issues and controversies as important and urgent as any

the economist has, if we are to make democracy work, or work better. Discovering economic truth comes first, but truth is of scant value unless the public and the government respect it and use it. And I believe that economists can do much more to improve economic understanding if they really give it the energy, resources, and thought the job requires.

The job can't be done adequately after people are adult and set in their ways; it needs to be started in the elementary schools. It needs to be done better through the mass media. And it needs to draw on the communications talents and imagination and enthusiasm of a larger number of students passing through our colleges and graduate schools. Most of all, it can be done more brilliantly and soundly, reaching more of the public, if the economics profession really means business and will provide the incentives and rewards for those willing to try.

Increasing the Public's Understanding of Economics: What Can We Expect From the Schools?

By MICHAEL A. MACDOWELL*

In answer to the question, "What can we expect of schools (K–12) in raising the level of economic understanding?," my first reaction is, "Not much." We can't expect much from the schools unless the economics profession is prepared to offer strong and continuing support that will advance the training of teachers and improve the quality of the materials available to help them teach economics. Therefore, my appeal in this paper is for increased assistance from the economics profession. The Joint Council of Economic Education (JCEE) has been in the forefront of the endeavor to teach economics at the primary and secondary school levels for some 37 years now. Lately, we have stepped up our efforts considerably, with encouraging results. Still, we have a long way to go, and we cannot broaden our reach as far as we should without more help from you and your colleagues.

Historically, too few teachers have had any training in economics whatsoever. According to the Southern Regional Education Board (1985), only 25 percent of graduating teachers' transcripts show even a single course in economics. The only other liberal arts subject in which they had fewer courses is philosophy.

Once in the classroom, teachers receive precious little in-service training to update their skills, according to William Walstead and Michael Watts (1985). Surveys of elementary teachers report that about half had no coursework, and another 25 percent had taken only one course. Surveys of secondary social studies teachers who specialize in teaching courses or units in economics show about 15 percent with no coursework and another 25 percent with one course; 30 percent reported taking only two courses in economics. In other words, 70 percent of teachers who teach economics have had two courses or less in the subject.

Further indication of the weak knowledge base of teachers in economics is demonstrated by the relative ranking they give to key concepts in the discipline. *The National Survey of Economic Education*, 1981, asked junior and senior high school teachers to rank economic concepts by their importance. The results show that 24 percent ranked the key concepts of tradeoffs (24 percent) and opportunity costs (34 percent) as relatively unimportant.

I. The Demand for Economic Education

Twenty-seven states now require some kind of economic education. Fifteen require a free-standing economics course, while the remainder recommend the "infusion" of key economic ideas at strategic points throughout the curriculum. In states where infusion was adopted, there is reason to believe that infusion is minimal. A study in New Hampshire which had an infusion clause until very recently, indicates that only 25 percent of secondary teachers and 10 percent of elementary teachers could define opportunity costs, economic systems, inflation, or deflation (Walstad and Watts, p. 141).

Teachers of the high school economics course have fared much better. Most have been teaching economics for some years, have greater background, and are more familiar with the materials and resources available. But there are few highly trained high school economics teachers. The vast majority teach economics only occasionally.

The biggest problem on the immediate horizon is to bolster the skills of teachers who have never taught the subject but *soon* will be required to do so. The increase in new

*President, Joint Council on Economic Education, 2 Park Avenue, New York, NY 10016.

state mandates is thrusting literally thousands of teachers into economics teaching without enhancing their capacity to do the job.

In Texas, 3,600 secondary social studies teachers are employed and could be called on to teach economics. Fewer that 5 percent (less that 180) have completed at least two courses in economics; only 1 percent (about 40) possess the equivalent of a minor in economics. In two other major states—New York and California—which require a free-standing economics course, teachers appear equally unprepared and readily admit it when asked.

Placed in the untenable position of having to teach an unfamiliar subject, some teachers and school districts lean on prepackaged courses that are purchased by local businesses for classroom use. These materials are not only expensive, but they stress subject matter such as business organization, marketing, and some basic accounting, which most economists would agree is tangential to the discipline. Anecdotal information indicates that many graduates of these one-semester high school economics classes haven't even heard of the "circular flow," much less understand it.

Given the plethora of new state mandates, we would expect state departments of education, school districts, and teachers themselves to react to the derived demand for their services by arranging for additional coursework in economics. But this isn't happening because the incentives are too few and too weak.

> Despite limited coursework and the feeling of inadequate preparation, the large majority of teachers do not want to take additional *courses* in economics, but prefer short workshops.... The limited interest in additional training is not due to the poor quality of inservice courses or workshops. Teachers generally rate the programs they have attended positively.
> [Watts and Walstad, p. 140]

The incentive problem is twofold. First, the most a teacher can expect from additional coursework in economics is another step upward on the district's salary scale. Only a handful of districts offer any additional financial inducements for coursework-mandated areas such as economics. Second, while twenty-seven states have mandated economics, only fourteen provide any funding to help cover the costs of preparing new materials and providing short-term workshops to aid their teachers.

II. The Role of the Profession

What should be the role of the economics profession? The American Historical Association back in 1899 established a "Committee of Seven," a group of respected historians who set forth a systematic four-year history curriculum that, until very recently, set a national pattern for the American high school. The American Economic Association, by contrast, did nothing for 60 years in addressing the issue of economic education in the nation's schools. Then, in 1961, it established the National Task Force on Economic Education in the nation's schools. An original member of that Task Force, G. L. Bach, summarized the economists' attitudes about economics at that time:

> The attitude of the economics profession towards economics in the high schools [all schools] has generally been one of disdain and disinterest...in general, economists have apparently felt (1) it is better to get beginning economics students in college who have not had high school economics, because you generally have to unteach a good deal of what they have learned in high school. (2) Economics is too difficult and too important to teach in the high school anyhow. (3) Hence it is better to do nothing at all with economics in the high school.... [1961, p. 580]

Has the situation changed much in the past 25 years? Yes and no. On the positive side, the American Economic Association's 1961 Task Force emphasized the importance of teaching economics in the schools, and sketched out what kinds of knowledge high school graduates should possess about economics. Shortly thereafter, the JCEE embarked on the Developmental Economic Education Program (DEEP) which estab-

lishes a contractual relationship between the JCEE and a local school district. DEEP has realized some substantial success. It currently includes 875 school districts enrolling about twelve million students. These DEEP teachers receive training in economics and in the use of JCEE materials. The program is serviced by a network of 50 state councils on economic education and 265 university centers for economic education, the majority of which are located in departments of economics.

By the mid-1970's, it became apparent that some reorientation was required. This led to the creation of the JCEE's *Master Curriculum Guide*. This *Guide* has spawned three major PBS television programs for students in the elementary and secondary grades, about 100 separate print and curriculum guides and materials for instruction, and a growing number of microcomputer modules for various grade levels. It serves as the basis for most of the new high school economics textbooks and a majority of the state mandates.

The JCEE and its network have expanded their teacher training efforts, reaching about 100,000 teachers last year, of whom at least 40 percent spent enough time on task to generate marked improvements in their students' learning. The JCEE's network and the AEA Committee on Economic Education have done a great deal to advance the cause of economic education in the schools.

The effects of these efforts over the past 25 years are what you might expect. In 1961, only 5 percent of students took economics prior to graduation from high school. Today, we can report that some 27 percent of all students are enrolled in school districts that are affiliated with the Joint Council on Economic Education. According to the national norming of the high school Test of Economic Literacy (1978), students in districts with a viable economics unit in their curriculum score 20 percent higher.

While we can count some progress since the release of the 1961 Task Force Report, we have also lost ground. In the first place, the cohorts of teachers with extensive knowledge of economics are close to retirement. In the second place, while we have been training more teachers, we are not training enough of them. We must both replace those who are retiring and expand the number of teachers equipped to meet the increase in state mandates for education in economics.

Perhaps the most troublesome sign is the waning interest of professional economists, and others who support economic education, in helping teachers teach economics effectively. The profession has never been, as Bach pointed out, captivated by economics education. Equally important, those economists who get involved have been less highly regarded than their collegiate peers.

Burgeoning undergraduate enrollments in economics have exacerbated the problem by reducing the number of special courses available for teachers in economics. This is occurring just when a distinguished task force, chaired by Henry J. Hermanowicz, Dean of Education at Penn State, and including James Tobin of Yale, is releasing new preservice guidelines for teacher education. These guidelines recommend that all teachers take at least one course in economics, and that secondary teachers take two to three courses. The task force goes on to recommend that, whenever possible, special training, or courses, be given just for preservice education majors. The intent is not to produce a watered down "principles" section for teachers. Rather it is to emphasize how economics can be applied and how available teaching materials can be utilized in the classroom.

III. Recommendations

It is imperative that the profession renew and re-energize its interest in economics education by 1) giving much stronger professional acceptance and recognition to their own colleagues who work in the field; 2) participating in programs to give economics instruction to those who will teach or are teaching the subject in primary and secondary schools; and 3) helping teachers to pick high-quality teaching materials and to reject those of dubious value or that promote special interests.

1) *Acceptance of Colleagues*. In many academic fields, professors help teachers at lower levels to understand the basics of their

disciplines. Across the field, economists do not seem to devote nearly as much effort to such "good works."

Major colleges and universities can change this trend by rewarding and acknowledging quality work in economics education, and thus encourage other institutions to do the same. And individual academic economists can join their distinguished colleagues who are already active in the efforts of the JCEE —G. L. Bach, W. Lee Hansen, Allen C. Kelley, William J. Baumol, Martin S. Feldstein, Rendigs Fels, Paul McCracken, Robert M. Solow, Herbert Stein, and James Tobin, to mention a few.

2) *Participating in Programs.* Professional economists are needed to establish and help staff centers for economic education. However, as a practical matter only a small percentage of the profession need get directly involved in economics education. At any one institution the long-term goal might be 5 percent of the staff time of major departments. For instance: a) One person, serving half-time, in most departments of ten to twelve economists; and b) One full-time (or full-time equivalent) in most departments of twenty or more.

3) *Materials Review.* In thousands of school districts, educators with no background in the subject whatsoever are forced to choose economic education materials. Schools badly need help in selecting quality textbooks and ancillary materials. Much of what is available and called economic education is self-serving and, far more important, does not begin to address the basic analytics of the discipline of economics. Professors will need to cooperate with state councils on economic education as well as state and local school boards and administrators to carry out the appraisal of materials.

III. Conclusion

While there are more students than ever before taking economics in the schools, teachers are still not adequately prepared, and the situation is not improving. Because of new mandates, the number of students taking economics is increasing at a time when the relatively few adequately prepared economics teachers are nearing retirement. Some progress was made by the profession in the 1960's and 1970's in support of economic education, but that support may be waning. Hopefully, the profession will rise to the occasion and renew its commitment to economic education to the schools. The time investment is minimal, and the potential for success great.

REFERENCES

Bach, G. L., "Economics in the High Schools: The Responsibility of the Profession," *American Economic Review Proceedings*, May 1961, *51*, 579–86.

Brennan, Dennis C., "A Survey of State Mandates for Economic Instruction," Joint Council on Economic Education, 1985.

Clark, J. R. and Barron, D. D., "Major Findings of the National Survey of Economic Education," *Journal of Economic Education*, Summer 1981, 45–51.

Ravitch, Diane, "Decline and Fall of Teaching History," *New York Times Magazine*, November 17, 1985, p. 101.

Saunders, Phillip et al., *Master Curriculum Guide in Economics: A Framework for Teaching the Basic Concepts*, 2nd ed., New York: Joint Council on Economic Education, 1984.

Wagner, Lewis E., "Task Force to Classroom," *American Economic Review Proceedings*, May 1963, *53*, 660–73.

Walstad, William and Watts, Michael, "Teaching Economics in the Schools: A Review of Survey Findings," *Journal of Economic Education*, Spring 1985, *16*, 135–45.

Southern Regional Education Board, "An Agenda for Higher Education and the Schools," Atlanta, 1985.

What Knowledge Is Most Worth Knowing—For Economics Majors?

By W. LEE HANSEN*

This essay resurrects an old question but asks it in a new context. The old question is a paraphrase of the title of Herbert Spencer's famous essay "What Knowledge is of Most Worth?" The new context is to pose the question for undergraduate students majoring in economics. My intent is to engage you in reflecting about what kinds of knowledge and skills our economics majors should master—what proficiencies they should be able to demonstrate—by the time they graduate from college. My focus is on not the select few who plan to enter graduate economics programs, but rather the vast majority who go out into the world and will become the next generation of leaders. I propose a list of knowledge and skills, perhaps a better word is proficiencies, that we might reasonably expect our majors to demonstrate upon graduation. This is by no means a final or definitive list; rather it is offered to stimulate discussion about the meaning of the economics major and how to give it more meaning.

This question, to the best of my knowledge, has not recently been raised in a serious way by the economics profession. It was touched on briefly more than 30 years ago in a special supplement to the *American Economic Review* on undergraduate education (see Horace Taylor, 1950). The silence since then is peculiar in light of significant professional interest in examining the outcomes of a wide variety of public programs through cost-benefit analysis, program analysis, and the like. It now seems timely to raise the question because of the growing concern about what a college degree represents, not in some grand philosophical sense but in readying young people for a fuller and more productive life (see the Association of American Colleges, 1985).

What do we expect of our graduating economics majors? We take as given their need to meet the requirements for graduation established by their colleges. Usually, this entails four years of course work beginning with general education courses in the first two years followed by completion of the major and other breadth requirements during the final two years. It is the major, however, that gives focus to the college experience. Typically, majors are required to take somewhere between 24 and 40 credit hours, or from 8 to 13 economics courses—elementary, intermediate, and advanced. These courses must usually include intermediate theory and statistics and perhaps also other distributional requirements (see John Siegfried and James Wilkinson, 1982). It goes without saying that students must earn passing grades in these courses.

What do our majors know after all this? About all we as members of economics department faculties can claim for our majors is their ability to answer our examination questions with some facility. Implicit in our approach is the assumption that exposure to a range of courses produces learning, learning that enables students not only to apply their knowledge to a variety of questions and issues they will confront as citizens but also to prepare them for employment as economists (albeit junior economists), or in related jobs that may utilize the skills acquired by economics majors. Whether graduating economics majors can in fact apply their knowledge as citizens or employees is not clear, to us and perhaps to them either.

Some economists might well dispute whether improved citizenship and job preparation are uppermost in the minds of eco-

*University of Wisconsin, Madison, WI 53706.

nomics majors or their instructors. They would assert that as members of economics departments we should not be concerned with such questions. The undergraduate degree is essentially a liberal arts degree that prepares students to think critically about the great and small issues of mankind; its purpose is not vocational; and most students end up doing work that is at most only loosely related to their undergraduate majors. Continuing, they would argue that the purpose of economics instruction is to offer an opportunity for students to be exposed to different facets of the discipline and to learn how economists go about their work. Responsibility for figuring out how to put together what is learned rests with students. How pervasive this view is remains to be ascertained. Whatever the case, economics is seen by many students as a discipline that can offer enlightenment about how the economics system operates as well as improved employment prospects after graduation. Indeed, numerous departments tout their programs as providing these opportunities.

While it may still be unclear who is responsible for what, this leaves unanswered the question of what we might hope or expect our majors to be able to do as a result of their heavy investment in schooling. Remember that they had extensive contact with the economics literature, the fundamentals of economic analysis, and applications to a variety of economic issues and problems. They have participated in somewhere between 300 and 600 hours of classroom instruction by professional economists. At the same time they have experienced a heavy diet of multiple choice examinations. Few have been subjected to any serious grilling in the classroom because of the dominance of lecturing. And not many have had to write papers in their courses. While we like to say that we have tried to teach them how economists think, we have almost no evidence as to our success. In fact, except for the exams they took, there may be little that distinguishes graduating economics majors from those in, say, biology or political science.

How do we specify what it is that economics majors should know? We can begin by examining statements of the competencies expected of college entrants. Perhaps the most comprehensive statement is that prepared by The College Board (1983) to describe for prospective college students what is most worth knowing. Less has been done at the college level though there is a strong move by some institutions to specify what students should have mastered in the general education phase—the first two years—of college. Statements about the general competencies expected of college graduates are usually quite vague. Even less has been done by academic disciplines to spell out how their instructional programs for majors enhance, or are intended to enhance, the capacities of students to demonstrate their knowledge and skills. Thus, the existing literature provides little or no guidance.

This gap offers a bold opportunity to set out a list of proficiencies for graduating economics majors. This list is based on thinking about the question for some years and on occasional efforts in my undergraduate teaching to build the competencies of my students. It also reflects the responses to surveys of our recently graduated majors, conversations with employers who hire our majors, and discussions with a range of other people—colleagues in economics, faculty members from other disciplines, and high-level executives in both business and government.

This list of proficiencies moves from what might be called lower to higher levels of cognitive activity, as suggested by the groupings of the items on the list.

1) *Gaining Access to Existing Knowledge*: Locate published research in economics and related fields; locate information on particular topics and issues in economics; search out economic data as well as information about the meaning of the data and how they are derived.

2) *Displaying Command of Existing Knowledge*: Summarize (in a 2-minute monologue or a 300-word written statement) what is known about the current condition of the economy; summarize the principal ideas of an eminent living economist; summarize a current controversy in the economics literature; state succinctly the dimensions of a

current economic policy issue; explain key economic concepts and describe how they can be used.

3) *Displaying Ability to Draw Out Existing Knowledge*: Write a precis of a published journal article; read and interpret a theoretical analysis, including simple mathematical derivations, reported in an economics journal article; read and interpret a quantiative analysis, including regression results, reported in an economics journal article; show what economic concepts and principles are used in economic analyses published in articles from daily newspapers and weekly newsmagazines.

4) *Utilizing Existing Knowledge to Explore Issues*: Prepare a written analysis (of say, 5 pages) of a current economic problem; prepare a decision memorandum (of say, 2 pages) for a superior that recommends some action on an economic decision faced by the organization.

5) *Creating New Knowledge*: Identify and formulate a question or series of questions about some economic issue that will facilitate investigation of the issue; prepare a 5-page proposal for a research project; complete a research study whose results are contained in a polished 20-page paper.

Several comments need to be made. First, these proficiencies probably strike most of us as being quite reasonable. Indeed, most if not all economics faculty members would hope that students, by the time they begin work in the major, already possess the generic proficiencies indicated here. This would allow faculty members to be concerned with sharpening these proficiencies within the context of economics. Second, these proficiencies are quite neutral with respect to the content of the economics major, except perhaps for students planning to enter graduate school. Thus, an emphasis on these proficiencies need not intrude on the intellectual focus of the major.

At the same time, problems arise. First, how can we determine whether students have acquired these proficiencies by the time they graduate; put another way, how would we test for these proficiencies? The obvious answer would be to devote the final several weeks of the senior year of college to a hands-on testing program that would permit students to demonstrate their proficiencies. Second, how might we structure the major and its related teaching program so as to provide reasonable assurance that students develop these proficiencies? In fact, relatively little might have to be done except to incorporate the acquisition of these proficiencies into individual courses or groups of courses. A senior research seminar would be helpful, however, in giving students experience with the development of new knowledge. Third, how do we enlist faculty support for a focus on the development of proficiencies among our economics majors? This is no doubt the toughest question to answer because many faculty members are likely to view a focus on proficiencies as simply another attempt to increase the costs of their undergraduate teaching, with no offsetting increase in benefits to them as faculty members.

Rather than attempting to resolve all of these questions here, it is probably best to think about whether the list of proficiencies presented above is reflective of the proficiencies we would like to see in our graduating majors, and how we might take some small steps to facilitate the acquisition of these proficiencies by our undergraduate majors.

Do these proficiencies reflect what we think is important? Even if we believe they do, what do recent and long-time graduates have to say about the appropriateness of these proficiencies? What do employers think about this approach and these proficiencies? Will these proficiencies enhance the employability of graduating seniors? Will they contribute to the career development of our seniors? If we agree on these questions, how do we go about answering them?

On the assumption that these proficiencies are viewed as important, how do we encourage their acquisition? It seems clear that economics faculty members are not going to rise up overnight and embrace the concept of proficiencies. This means we have to find ways of making it easier for individual faculty members to promote the notion of competencies in their own teaching. What can be done? One approach is to develop materials that can be helpful to instructors.

Of particular value would be sample assignments and evaluations of actual student responses, both to show how to implement a proficiency approach and how to evaluate student achievement of these proficiencies. Another approach is to specify how the cumulative acquisition of these proficiencies can be assured. This requires development of a sequence of materials that would be integrated across courses in the major. None of these tasks is easy; indeed, implementation is the most difficult of all.

In developing these materials and integrating them into the major, it is important to assess whether the stress on proficiencies will subtract from the economic content of the major. It seems probable that adoption of the proficiency approach will lead to some restructuring of our teaching. The likely result is that students will be taught less, but will learn more, and learn what they do learn better than they do now. The impact on faculty members is more difficult to speculate about, but they could find this approach a stimulating one.

To sum up, if we are concerned about the public's understanding of economics, and everyone judges it to be low, perhaps we should concentrate on increasing the capacity of those students with whom we have the greatest contact, our majors. The proficiency approach outlined here should do at least as well as we now do and probably much better.

REFERENCES

Siegfried, John J. and Wilkinson, James T., "The Economics Curriculum in the United States: 1980," *American Economic Review Proceedings*, May 1982, *72*, 125–38.

Taylor, Horace, "The Teaching of Undergraduate Economics: Report of the Committee on the Undergraduate Teaching of Economics and Training of Economists," *American Economic Review*, December 1950, *40*, Suppl., Part 2.

Association of American Colleges, *Integrity in the College Curriculum: A Report to the Academic Community*, Washington, D.C., February 1985.

The College Board, *Academic Preparation For College: What Students Need To Know And Be Able to Do*, New York, 1983.

Public Policy and Productivity in the Trucking Industry: Some Evidence on the Effects of Highway Investments, Deregulation, and the 55 MPH Speed Limit

By THEODORE E. KEELER*

The past 15 years have seen important changes in public policies toward the trucking industry, including deregulation (the Motor Carrier Act of 1980), the imposition of a nationwide 55 mile per hour (mph) speed limit, and significant changes in the stock of capital invested in the nation's highway system. The last of these changes entailed a substantial increase in the nation's highway stock in the 1960's through the mid-1970's. Since that time, however, it is widely believed that the system has not been kept up to earlier standards, with a resulting decline in highway capital stock likely.

The main purpose of this paper is to analyze the effects of the last two of these policy changes on productivity in trucking (speed limits and infrastructure changes) over the last twenty years. It will also provide some tentative evidence of the effects of deregulation on costs for some large trucking companies.

Although it is possible to suggest the directions of these effects a priori, it is impossible to do so with certainty. In the case of the 55 mph speed limit, faster speeds mean better utilization rates of the trucks and drivers, but they also mean higher fuel consumption, and, if the trucks are not engineered to operate at very high speeds, they can mean higher operating costs as well, so the direction of the effect is ambiguous.

In the case of infrastructure, higher investments should, all other things equal, make for lower costs. But a connection between infrastructure investment and productivity has been difficult to establish through econometric measurements of other parts of the economy (C. R. Hulton and R. M. Schwab, 1984).

Finally, one should expect truck deregulation to lower costs: entry of or expansion by new, nonunion firms, combined with deregulated rates, has forced many carriers to negotiate lower wages. And the operating flexibility achieved by elimination of route restrictions has allowed many carriers to reduce empty backhauls, thereby lowering costs. Here, too, however, one must be careful in making one's predictions. It may be true that deregulation will reduce trucking costs through lower factor prices. But, once one controls for factor prices (as is done in a cost function), the effect is no longer so unambiguous. It may be that with the lower factor prices afforded by reorganization of the industry, many shippers would prefer to take some of the benefits of lower costs in the form of improved service. If that is true, with factor prices controlled for, it is possible that truck deregulation could cause costs to rise. In the case of all three hypotheses tested then, there is some a priori reason to be cautious in predicting which direction the variable will take on observed trucking productivity.

The method used for testing the hypotheses is a translog cost function, estimated with pooled cross-section and time-series data for a panel of twelve large, interregional trucking companies over the period 1966–83.

†*Discussants*: Jose A. Gomez-Ibanez, Harvard University; Gregory K. Ingram, World Bank.

*Department of Economics, University of California, Berkeley, CA 94720. I am grateful for the support of the Institute of Transportation Studies of the University of California and the research assistance of John Ying.

I. The Model

I assume that the productivity effects we wish to measure manifest themselves through costs. To the extent that the cost function does not control for service quality, and to the extent that infrastructure improves service quality, this approach will understate the productivity effects of the changes.

The cost function estimated here is of the translog form, and is in some ways similar to cost functions estimated by A. F. Friedlaender and S. J. W. Chiang (1983) and A. F. Daughety, et al. (1986). It represents a general polynomial approximation in logs for a function including the following variables: q, an output variable; $(w_1,\ldots,w_4)$, a vector of four factor prices (discussed below); $(z_1,\ldots,z_3)$, a vector of nonoutput firm-attribute variables (including average length of haul, shipment size, and a highway infrastructure variable for the areas served by a particular firm at a particular time); *D55*, a variable allowing for the presence of the 55 mph speed limit; *Dereg*, a variable accounting for deregulation; and T, a time trend variable, picking up changes in productivity not otherwise accounted for.

The specific form of the cost function, as estimated, is as follows (all variables except the dummies are divided by their sample means):

$$(1)\quad \ln TC = A + B\ln q + \sum_i C_i \ln w_i + \sum_j D_j \ln z_j + E\ln T + FD55 + GDereg + \sum_n H_n DF + 1/2\sum_i\sum_l J_{il}\ln w_i \ln w_l + \sum_i\sum_j K_{ij}\ln w_i \ln z_j + \sum_i L_i \ln w_i \ln q + \sum_i M_i \ln w_i \ln T + 1/2\sum_j\sum_m N_{jm}\ln z_j \ln z_m + \sum_j P_j \ln z_j \ln q + \sum_j R_j \ln z_j \ln T + 1/2V(\ln q)^2 + W\ln q\ln T + 1/2Y(\ln T)^2 + e,$$

where e is an econometric error term, $J_{il} = J_{li}$, $N_{jm} = N_{mj}$, for all i, j, l, and m, and the DF are firm dummy variables whose purpose is explained below.

Economic theory also requires the following constraints to assure that the function is homogeneous of degree one in factor prices: $\Sigma_i C_i = 1$; $\Sigma_i J_{il} = \Sigma_l J_{il} = \Sigma_i \Sigma_l J_{il} = 0$ for all i, l; $\Sigma_i K_{ij} = 0$ for all j, $\Sigma_i L_i = 0$; $\Sigma_i M_i = 0$.

This cost equation was estimated jointly with appropriate factor share equations (as done by L. R. Christensen and W. H. Greene, 1976), which, following Shephard's Lemma, is derived by differentiating the cost equation with respect to the log of each factor price. Thus, if Si is the share of factor i in total costs and ε_i is an econometric error term for each share equation, we have

$$(2)\quad \partial \ln TC/\partial \ln w_i = S_i = C_i + \sum_l J_{il}\ln w_l + \sum_j K_{ij}\ln z_j + L_i \ln q + M_i \ln T + \varepsilon_i.$$

For n factors (four in this case, as explained below), only $n-1$ of these equations are independent, given that factor shares add up to one.

The cost equation affords the possibility of testing numerous hypotheses about costs and production in the trucking industry (which must be reported in another paper, due to space constraints). My goal here is to test the above-mentioned hypotheses about productivity, infrastructure, the 55 mph speed limit, and deregulation. These hypotheses should be testable on the basis of the signs of the coefficients representing each relevant variable. Thus, if the infrastructure variable increases with more highway investment in a given firm's territory, then that coefficient ought to have a negative sign in the cost equation, if in fact infrastructure investments have decreased costs. If the 55 mph speed limit reduced costs, the dummy variable specified would have a negative sign, as would the deregulation variable if it reduced costs.

In addition to the variables discussed so far, the estimated cost equation contains dummy variables for each firm in the sample save one. As pointed out in previous studies, by controlling for firm effects, these variables

should make for efficient estimates of the remaining parameters; see M. Balestra and M. Nerlove (1966) and D. W. Caves et al. (1984).

II. The Data Sample

The sample for this study was a panel of twelve firms with data over the 1966–83 period. The firms include Carolina Freight, Consolidated Freightways, Gateway, Leeway, McLean, Overnite, Pacific Intermountain Express, Roadway, Smith's, TIME-DC, Transcon Lines, and Yellow Freight. Because the data were not available for some firms for some years, or because some of the firms changed so dramatically in size that their data for some years were deemed inappropriate for inclusion, we were left with 204 observations. Data sources are listed in the footnote to Table 1.

Following previous studies, the output variable selected was ton-miles of freight (RT). Operating characteristics include average length of haul (AL) and average shipment size (SS). I have tried to include more detailed variables on type of freight hauled, since the carriers selected represent a rather homogeneous group of long-haul carriers of general commodities.)

Unlike previous studies, firms' capital costs are based on real, as opposed to nominal, values of capital stock. Those values were calculated by adding up gross investments (corrected by a truck capital goods price index) over 10 previous years, using the "one-horse-shay" or annuity depreciation assumption. This approach is consistent with that recommended by M. J. Peck and J. R. Meyer (1965) in transportation, and its continuing appropriateness for the U.S. economy is pointed out by M. J. Feldstein and L. Summers (1977). Book value of gross investment in a given year is measured as change in book value of net stock from the previous year plus depreciation in the given year.

I have assumed four separate factors, including labor (for which the factor price PL is measured by compensation per employee); capital (whose factor price PK is measured by the capital goods price deflator for trucks), along with the assumption of a cost of capital of 14 percent over the period (consistent with Friedlaender and Chiang); fuel (for which the price PF is a straightforward price index for diesel fuel); and materials and other inputs (price PO is measured as total other expenditures per vehicle-mile).

Unlike some previous studies of truck costs, I do not include purchased transportation (trucks rented from other carriers) as a separate factor. Rather, purchased transportation expenditures have been aggregated with capital costs, for two reasons. First, although some trucks are rented with drivers, most are not, and so truck rentals can best be viewed as capital expenses. This is consistent with Friedlaender and Chiang's finding of a very high elasticity of substitution between capital and purchased transportation. Second, and more importantly, accounting practices of some truck firms over the time period studied made disaggregation of these two factors a serious problem. Specifically, for some years of the period, some firms formed common carrier subsidiaries which leased their trucks from the parent company, apparently for tax reasons. Thus, the asset values of the common carrier firm jumped around considerably, as did purchase transportation expenses, the asset values declining as the purchased transportation costs rose, and vice versa. Meanwhile, assets of the parent firm (as listed in *Moody's Transportation Manual*) were far more stable. The aggregation of capital costs and purchased transportation expenses resulted in a far more stable (and arguably realistic) estimate of total capital costs.

To calculate highway infrastructure (HC), capital stocks for all state-maintained highways were calculated for each of the 48 contiguous states for the year 1955 and for each of the years 1966–83, using data from the U.S. Federal Highway Administration (FHWA). These stocks, once again, are based on real values, calculated with the annuity form of depreciation and the FHWA's highway construction cost index, assuming a 25-year average lifetime for highways (this lifetime is obviously too short for bridges and too long for pavements; but on the average, it seems about right; see my paper with K. A. Small, 1977). Once these stocks were calculated, they were assigned to each carrier in each year by looking at a map of the states

covered by that carrier. Thus, an aggregate stock was calculated for each sample point based on what states the carrier served. Finally, the total stock of highways so calculated for each carrier was divided by the appropriate 1955 stock for the same states (i.e., the states actually served by each carrier in the appropriate year). The year 1955 was selected as being appropriate for a deflator because it just precedes the beginning of the U.S. interstate highway system and the large infrastructure buildup which it (and other limited-access highways) entailed. On the other hand, as of 1955, most places were served by paved, two-lane roads. The variable for each carrier thus reflects the extent that the territory served by that carrier in a given year has been built up beyond the basic 1955 system.

This last step of dividing highway capital stocks by 1955 values was necessary to control for the fact that some states are bigger than others, and a bigger state does not automatically have better infrastructure. The variable I have constructed should represent an accurate value of the quality of the infrastructure for the territory served by each trucking system.

Finally, reflecting the 55 mph speed limit is a dummy *D55* valued at 0 before 1974, at 1 after 1974, and at .5 during 1974. Similarly, the deregulation variable *DEREG* has the value of 0 before 1981 and 1 thereafter.

TABLE 1—REGRESSION RESULTS

Variable	Coefficient	S.E.	Variable	Coefficient	S.E.
Constant	−0.0679	0.0315	*AL·PK*	−0.0069	0.0058
AL	−0.4645	0.0396	*AL·PL*	−0.0384	0.0052
RT	1.0345	0.0268	*AL·PF*	0.0123	0.0016
SS	−0.4662	0.0422	*AL·HC*	−0.0305	0.3086
PK	0.1534	0.0027	*AL·T*	0.0682	0.0613
PL	0.5904	0.0024	*AL·PO*	0.0331	0.0029
PF	0.0732	0.0008	*RT·SS*	−0.1483	0.0447
PO	0.1830	0.0014	*RT·PK*	−0.0011	0.0036
HC	0.1996	0.2270	*RT·PL*	−0.0009	0.0032
T	−0.0086	0.0533	*RT·PF*	−0.0005	0.0010
D55	−0.0357	0.0178	*RT·PO*	0.0024	0.0019
DEREG	0.0649	0.0173	*RT·HC*	−0.1351	0.2011
AL·AL	0.1247	0.0784	*RT·T*	−0.0456	0.0383
RT·RT	0.1283	0.0380	*SS·PK*	0.0674	0.0070
SS·SS	0.4561	0.1167	*SS·PL*	−0.0804	0.0063
PK·PK	−0.0035	0.0220	*SS·PF*	−0.0036	0.0021
PL·PL	0.1672	0.0145	*SS·PO*	0.0167	0.0037
PF·PF	0.0395	0.0038	*SS·HC*	0.0587	0.4960
PO·PO	0.1211	0.0051	*SS·T*	0.0365	0.0941
HC·HC	6.8545	4.3780	*PK·PL*	−0.0301	0.0154
T·T	0.2004	0.1848	*PK·PF*	0.0234	0.0067
AL·RT	−0.0796	0.0367	*PK·PO*	0.0102	0.0080
AL·SS	0.2398	0.0577	*PK·HC*	−0.0801	0.0540
PK·T	0.0150	0.0113	*TRANS-CON*	−0.0109	0.0436
PL·PF	−0.0344	0.0043	*MCLEAN*	0.0182	0.0343
PL·PO	−0.1028	0.0057	*SMITHS*	0.0228	0.0229
PL·HC	−0.0308	0.0485	*CAROLINA*	0.0325	0.0240
PL·T	0.0127	0.0101	*OVERNITE*	0.0185	0.0419
PF·PO	−0.0285	0.0029	*ROADWAY*	−0.2136	0.0527
PF·HC	0.0833	0.0155	*CONSOLI-DATED*	−0.1987	0.0534
PF·T	−0.0170	0.0034	*LEEWAY*	−0.0300	0.0346
PO·HC	0.0276	0.0288	*PIE*	−0.0035	0.0318
PO·T	−0.0108	0.0060	*TIMEDC*	−0.0021	0.0396
HC·T	−1.1750	0.8959	*YELLOW*	−0.0801	0.0473

R^2 for cost equation = .9939   *RMSE* = .00451
R^2 for *k* share equation = .4686   *RMSE* = .00059
R^2 for *l* share equation = .7667   *RMSE* = .00046
R^2 for *f* share equation = .7928   *RMSE* = .00005

Source: Data for regressions are calculated from the following sources (see text): Trinc Transportation Consultants, *Trinc's Blue Book of The Trucking Industry*, 1967–84 and U.S. FHWA, Highway Statistics, 1930–83. In certain cases in which accounting data were not available from *Trinc's*, they were taken from *Moody's Transportation Manual*, various years.

III. Estimation and Results

The cost equation described in (1) was estimated using the maximum likelihood interation of A. Zellner's (1962) method of seemingly unrelated coefficients, and the results are shown in Table 1. Because all variables were specified as deviations around means, the coefficients of the first-order terms represent point estimates of the elasticity of costs with respect to each variable with all variables held at their means.

Before discussing the results for the coefficients of interest regarding speed limit, infrastructure, and deregulation, I consider very briefly estimates of other parameters, to get a sense of the overall plausibility of the cost function. The coefficient for *RT* is very near one (slightly above it, though insignificantly so) so it would seem consistent with constant returns to traffic density and scale (it is worth noting that the scale economies coefficient is very near that of Daughety et al., though they are able to reject constant returns, whereas I am not). As expected, a longer haul or a larger shipment size has a strong negative impact on costs. The factor prices all have cost elasticities which seem appropriate, and they are highly significant. Finally, the negative first-order coefficient for *T* indicates that at the grand sample mean, productivity (unexplained by other variables) did rise, although the coefficient is not statistically significant.

I now turn to analyzing the effects of the three variables of interest. First, the 55 mph speed limit has a negative and significant effect on costs. This suggests that the American Trucking Association was following its own best interests in failing to oppose this law, at least as it applies to the large trucking firms in my sample. Nevertheless, the improved service afforded by higher speeds might be worth the costs, so these results do not necessarily support the current speed limit as a policy.

Second, the infrastructure variable is positive but insignificant. This suggests that at my sample mean, an incremental dollar's worth of infrastructure investment had no meaningful impact on costs. There are several possible reasons for this surprising result. First, other econometric effects to measure the productivity effects of infrastructure (in areas such as manufacturing) have been no more successful than the present one. Second, the better infrastructure available at some times and in some regions may show up in the form of better truck service quality, rather than lower costs. Third, it may be that because so many of the actual capital investments in our highways over recent years have been targeted on the margin to help autos, measuring effects for trucks may be difficult.

Finally, the results on truck deregulation are unexpected, in that they indicate that reform had a significantly positive effect on costs for the firms in our sample. There are several possible reasons for this. First, my estimate is of the effect of deregulation on costs *controlling for factor prices*. It is likely that with the lower labor costs occurring with deregulation, shippers would prefer some of the benefits in the form of better service, rather than lower rates, in turn causing costs to rise, holding factor prices constant. Furthermore, some firms in my sample may have had difficulty adjusting to deregulation, so as a result they lost business and their productivity declined because factors were not adjusted appropriately. Finally, the 1982–83 recession could have been responsible for the postderegulation productivity decline. My conclusion in this area is therefore only suggestive—further study is needed as more data become available.

Because the dummy variables for deregulation and speed limit do not contain higher-order polynomial terms, the results for them are valid throughout the sample. In the case of the highway infrastructure variable, the results reported here apply only with all variables at sample mean values. Although simulations done for other values do not seem to change my qualitative results, space limitations preclude reporting them in this paper.

REFERENCES

Balestra, P., and Nerlove, M., "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, July 1966, *34*, 585–612.

Caves, D. W., Christensen, L. R. and Tretheway, M. W., "Economies of Density versus Economies of Scale: Why Trunk and Local Service Airline Costs Differ," *Rand Journal of Economics*, Winter 1984, *15*, 471–89.

Christensen, L. R. and Greene, W. H., "Economies of Scale in U.S. Electric Power Generation," *Journal of Political Economy*, August 1976, *84*, 655–76.

Daughety, A. F., Nelson, F. D. and Vigdor, W. R., "An Econometric Analysis of the Cost and Production Structure of the Trucking Industry," in A. F. Daughety, ed., *Analytical Studies in Transportation Economics*, Cambridge, Cambridge University Press, forthcoming 1986.

Feldstein, M. J., and Summers, L., "Is the Rate of Profit Falling?," *Brookings Papers on Economic Activity*, 1:1977, 211–27.

Friedlaender, A. F. and S. J. W. Chiang, "Productivity Growth in the Regulated Trucking Industry," *Research in Transportation Economics*, Vol. 1, 1983.

Hulton, C. R. and Schwab, R. M., "Regional Productivity Growth in U.S. Manufacturing: 1951–78," *American Economic Review*, March 1984, *74*, 152–62.

Keeler, T. E. and Small, K. A., "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways," *Journal of Political Economy*, February 1977, *85*, 1–25.

Urban Road Reinvestment: The Effects of External Aid

By GEORGE E. PETERSON*

There is a widespread conviction that the United States should be doing more to repair and improve its roads. The other papers in this session examine investment choices affecting the interstate highway system. In this paper, I look at the opposite extreme of the nation's road network: the local road system under urban government administration.

The deterioration of city streets probably is linked as closely with a general under-maintenance of urban capital facilities, and postponement by cities of capital reinvestment decisions of all kinds, as it is to characteristics of the transportation sector. Cities with backlogs of street repairs are likely to confront backlogs for water pipe replacement, separation of combined sewer systems and sewer pipe replacement, and bridge and building repairs, as well.

In 1982 when it passed the Surface Transportation Assistance Act, Congress recognized the special problems of city road repairs. It preserved the Federal Aid–Urban Program, which the administration had sought to eliminate, and for the first time required that 40 percent of the monies be used for rehabilitation, resurfacing, repair, or reconstruction, starting in 1984. A handful of states have also created special grant programs to assist cities in resurfacing city streets. In addition, capital repairs have acquired local political impetus in some cities. A favorite device has been the establishment of community reinvestment efforts, that bring together business leaders and political leaders to form infrastructure commissions of several years' duration, whose purpose is to win local public support for higher capital investment and maintenance levels.

Have these efforts succeeded in reversing the trend toward postponement of urban road maintenance and repair? If so, which instruments of support have been most successful and hold most promise for the future?

I. Past Studies

If there is a conventional wisdom among current analysts, it is that external aid programs like Federal Aid–Urban (FAU) have little effect in stimulating local expenditures on targeted activities. Even when they promote additional spending, much of the extra expenditure is thought to be dissipated in higher costs rather than translated into enhanced service capacity.

The last 15 years of estimates of the impact of grants-in-aid on state-local spending have brought successively lower estimates of expenditure coefficients. Recent studies of the stimulative effect of unrestricted grants have fallen to the point where, as theory would suggest, such grants appear to stimulate no more spending than a comparable amount of income in local taxpayers' hands. Edward Gramlich (1978, 1982) found that each dollar of federal revenue sharing resulted in only a few cents of additional state-local expenditure.

Even categorical matching grants have been found to be less stimulative than previously estimated. One reason is that most categorical grants are now closed-ended; that is, they are subject to fixed ceilings on the total amount of grant outlay. Such grants match local expenditures—and thus have price effects at the margin—only up to the maximum grant amount. Beyond that point, the grants carry no price subsidy. If a state or local government is already spending more than enough to qualify for the maximum federal grant, it is operating in a zone where the grant structure does not affect any of its incremental expenditure decisions.

Federal and state highway aid is a good example of a capped assistance program. In an analysis of the impact of federal aid

*The Urban Institute, 2100 M Street, NW, Washington, D.C. 20037.

on spending for primary, secondary, and urban highways (excluding interstates), Harry Meyers (1985) finds that each dollar of federal program assistance produces $0.37 of increased state spending on the designated target. The rest of the federal funding is spread into other nominally unaided highway activities. Although the matching structure of the federal grant program appears to be highly stimulative of state spending (offering a 3-to-1 match for state expenditures), thirty states in 1982 spent more of their own funds than the maximum that could be matched under the federal grant program. These states made expenditure choices at the margin based on the full cost of highway improvements. Meyers' analysis is consistent with earlier studies (Edward Miller, 1984) concluding that the *ABC* highway program had a small stimulative effect on state spending. Miller used data from the 1960's, when all but nine states spent more of their own funds than were eligible for federal matching.

Some qualitative examinations of the effects of the FAU program have judged there to be still less spending effect. The Congressional Budget Office (1982), for example, concluded that "the urban program has become essentially a form of revenue sharing," whose funding is almost fully fungible with local monies and therefore has minimal impact on expenditures for any specific purpose. The General Accounting Office (1984) interviewed transportation officials in seven states. These officials reported that they had not had any difficulties in meeting the 1978 Highway Act requirement that 20 percent of a state's noninterstate primary and secondary highway funds be used for 3*R* work. They did not anticipate difficulties in meeting the 1982 Act's requirements that states spend 40 percent of primary, secondary, and urban funds on 4*R* work, and felt that such requirements were unnecessary since states would be directing their highway investments to these purposes in any event.

Even when federal matching grants succeed in inducing higher state spending for the targeted purpose, there is evidence for some capital assistance programs that much of the higher outlay is captured in higher real costs of service provision. In its study of wastewater treatment grants, the Congressional Budget Office (1985) found that the capital costs of treatment plants were highly sensitive to the local share of project costs. Plants assisted by federal dollars tended to incorporate more costly design features, have excessive reserve capacity, and have their construction costs supervised less vigorously. For some sizes of plants, a decrease in the external funding share from 75 to 55 percent (as has been incorporated in recent federal aid legislation) was estimated to imply no increase in locally borne costs. The efficiency gains from having a larger local cost share fully offset the costs shifted from the federal government to local government by cutting the federal matching rate. Early studies of federal capital grants for urban bus replacement estimated that more than 20 percent of the value of the federal grant could be dissipated through inefficient substitution of new bus purchases for bus maintenance, if local transit authorities responded rationally to the federal subsidy structure (William Tye, 1973).

II. The Sample and Data

My concern is with local road preservation, which is best measured in terms of pavement condition or remaining pavement life. However, the techniques that cities employ to prolong pavement life run the gamut from seal coating through resurfacing of various depths (typically 1 to 3″) to total reconstruction. To combine these resurfacing efforts into a single index of road preservation requires taking into account the greatly different additions to design life that result from each technique.

The Urban Institute has conducted two surveys of local street resurfacing (Peterson et al., 1984; Peterson and Parham, 1985). In each, cities were asked to identify, by year, the miles of pavement that were seal-coated, resurfaced, reconstructed, or otherwise improved, by category of surface improvement. Resurfacing materials and depths of overlay were recorded. A design-life extension was assigned to each type of improvement, based upon local engineering estimates where these

were commonly used and upon standard engineering estimates for the region, if local estimates were not regularly employed. The sum of the "mile-years" of additions to design life of the local road network then was divided by the total number of road miles in the local inventory to compute an annual replacement rate, based on the assumption that each mile of local roadway, without repair, would lose one year of useful life with each year of aging. (This index borrows from Roy Eckrose, 1979.) For example, a city that replaces 50 miles of pavement with a design life of 20 years has added 1,000 mile-years to its inventory. If it has a total inventory of 1,500 miles, this would constitute an annual replacement rate of 67 percent of the useful life assumed to be lost through annual depreciation. Surveys were conducted in 1981 for the years 1978–80 and in 1985 for the years 1981–84.

An index of this type obviously is an overly simplified measure of pavement maintenance. It ignores the importance of traffic density and weight in road depreciation, except for recognizing differences in the extension of useful life that the same resurfacing techniques produce when applied to residential or arterial streets. It ignores the nonlinearity of actual road deterioration. (As long as a city's street inventory is evenly distributed by age, the assumption of linear depreciation is less distorting in the aggregate than it would be for individual road segments.) The estimates of design life extensions for different types of resurfacing were differentiated by road type and region, but probably should be considered no more than rough engineering guidelines. Only a few cities have begun to keep accurate records of the service lives produced by different resurfacing methods under actual local conditions.

Nonetheless, the variations between cities in pavement maintenance, and the variations within cities over time, are so large as to minimize the importance of measurement error. Table 1 shows pavement replacement rates for a sample of cities over the period 1978–80. The 40-city averages for 1978 are 63.7; for 1979, 55.8; for 1980, 50.6; for 1981, 50.3; for 1982, 48.9; for 1983, 51.5; and for

TABLE 1—STREET RESURFACING EFFECTIVENESS (Average 1978–80)

City	Score	City	Score
Chicago	1	Rochester	60
Detroit	6	St. Louis	61
Oakland	7	San Jose	61
Buffalo	12	Baltimore	63
Cleveland	13	Tucson	65
New Orleans	14	Phoenix	75
Houston	16	Anchorage	75
Albany	19	St. Paul	79
Shreveport	21	Sioux Falls	83
Philadelphia	22	Kansas City (MO)	87
Seattle	26	Green Bay	89
Miami	29	Denver	89
Washington, D.C.	31	Dallas	96
Lincoln	35	Charlotte	91
San Diego	41	Portland	91
Ogden	44	Milwaukee	112
Lexington	44	Atlanta	120
Wilmington	51	Garden Grove	121
Louisville	52	Cincinnati	159
Billings	54		

Source: The Urban Institute, from data provided by cities. Scores measure percentage of annual road depreciation replaced through resurfacing.

1984, 62.6. These figures seem to give support to the perception of undermaintenance of local streets. For the sample as a whole, between one-half and two-thirds of the assumed loss of useful life of the existing road inventory was replaced each year through resurfacing of reconstruction. Moreover, the replacement rate declined steadily for the first 5 years of the survey, before turning upward in 1983 and improving quite sharply in 1984. It should be noted, though, that the figures measure replacement of surfaces on existing roadways only. Construction of new road segments and widening of existing segments are not taken into account. Therefore, it is possible that the total inventory of real street assets was being sustained, or even added to, despite our evidence of only partial surface replacement.

III. Analysis of Local Street Resurfacing Decisions

We can contrast two models of local decision making about road resurfacing. One is a model that treats road resurfacing as a public product whose output level, like that of other public services, is determined by taxpayer demand, based on economic vari-

ables. However indirectly these demands are expressed through the local political system, they may be compressed into a demand function that expresses road resurfacing as a function of community income, the relative price of road construction, and external grant assistance.

The alternative model treats road resurfacing more as a budget-allocation decision made within the institutions of urban government. Local public officials, for example, often explain spending on street preservation as a residual. In years when a budget surplus is available, much of it is channeled into street repairs or resurfacing. When a city government faces an unexpected budget deficit, it is likely to close the gap through postponement of ordinary road improvements. In fact, the cities that have reacted most strongly to the publicity surrounding the urban infrastructure "crisis" have given first priority to breaking this cycle, wherein maintenance and repairs are determined as a residual after other claims have been met from the budget. Some cities (for example, Cleveland, Seattle, San Jose), have set up blue-ribbon infrastructure commissions to establish maintenance and reinvestment objectives for the city and to bring the pressure of the business community to bear on meeting maintenance goals. In some city-manager cities (for example, Dallas, Cincinnati), the professional staffs sometimes have placed maintenance and repair programs at the top of the city administration's management goals. Elsewhere (for example, Minnesota, Georgia), the state government has stepped in with new programs of assistance, restricted expressly to financing local road repairs.

There is very little support in my sample for a traditional demand specification of road resurfacing activity. When specified as a log-linear demand function fit to the 280 pooled cross-section and time-series observations, only the income variable was significant and of the right sign. The income coefficient implied a low-income elasticity of demand (less than 0.1) for local road repairs. The price variable, measured as the ratio of state highway construction costs to the general price level, was of the wrong (positive) sign. Road repair costs are an endogenous variable, strongly influenced by the level of repair and construction activity. A city's road repair expenditures tend to move in line with the state's total road spending and with national road spending. Road construction costs are highly sensitive to the overall volume of road construction, with the paradoxical result (from the perspective of a demand equation) that real purchases of local highway improvements and road construction prices move in tandem with each other.

Within the context of this poorly fit demand equation, the external aid variable is at the margin of statistical significance. A given volume of external aid (FAU or state aid for local road repairs) has roughly five times the impact on road resurfacing that the same amount of community income has. However, the coefficient on external assistance is very imprecisely estimated. It is perhaps unreasonable to expect a closer estimate, since there is little variation in FAU assistance over the sample. Federal funding allocations are based on a fixed formula, that assigns all urban areas roughly the same per capita assistance. Over my observation period, there were only moderate variations in the aggregate level of FAU funding (which declined from \$993 million in 1980 to \$800 million in 1983 and 1984). Consequently, most of the variation in this variable comes from the handful of state programs that channel assistance to local governments.

The large differences in local resurfacing activity are much better explained when institutional and budgetary variables are introduced into the expenditure equation. The local budget surplus or deficit position exerts a strong influence on repair activity. In a specification that uses spending, rather than miles added to surface inventory life, as the dependent variable, local road repair outlays climb more than 15 cents for each dollar of additional surplus available at the close of the previous fiscal year, and fall by a comparable amount for each dollar of local deficit. In some cities, this relationship has been formalized. The Seattle City Council in 1982 authorized that 25 percent of each year's unappropriated, unencumbered year-end balance in the general fund should be deposited in a repair and preservation fund to

finance major maintenance work. The Dallas City Council has, since 1982, formally dedicated a proportion of each year's closing budget surplus to the city's maintenance and repair backlog. In most cities, however, the process is less formal than this.

Dummy variables introduced to isolate city-manager cities and cities that have set up blue-ribbon infrastructure commissions to advise on maintenance repair policy also were found to be highly significant. The eight cities in the sample that had established formal public-private infrastructure policy commissions, with mayoral endorsement, carried out 30 percent more resurfacing, once these commissions were in operation, than they otherwise would be estimated to do. City-manager cities completed significantly more repair work than either "strong-mayor" or city council cities.

The importance of these institutional variables suggests that estimates of the effect of external aid should also take them into account. Table 1 makes clear that there is a group of cities for which the road resurfacing backlog is particularly severe. These tend to be cities that began incurring operating deficits in the mid-1970's and have not fully recovered from them. For these cities, the provision in the Surface Transportation Assistance Act requiring that, from 1984 onward, 40 percent of FAU funds be used for repair, rehabilitation, resurfacing, or reconstruction may well be a binding constraint. State resurfacing assistance may also require levels of spending that exceed the local pre-assistance level.

To allow for differences in the cities, I partitioned the sample into two groups: cities that ran up a cumulative budget deficit or faced back-to-back current deficits at some point over the course of the sample years, and those cities that did not. Resurfacing activity and road repair spending were then estimated separately for the two groups.

For the second group of cities—those without special budgetary difficulties—the spending requirements for 4*R* work under FAU, starting in 1984, and any similar requirements attached to state-aid programs had little impact on either 4*R* spending or physical resurfacing levels. The grant coefficient was less than in the pure demand specification (about 25 cents of stimulus per program dollar).

For the other "distressed" cities, however, the dollar requirements for 4*R* spending became powerful influences. Almost 80 cents of each dollar of aid within the minimum spending requirement has been translated into higher local road repair expenditures. Several of the distressed cities in our sample boosted road resurfacing dramatically in 1984, apparently in part to meet the requirements of FAU.

For the entire universe of local governments, federal aid for road resurfacing may have little expenditure impact and federal requirements for minimum resurfacing activity may be redundant. However, for the class of cities that have postponed maintenance and repairs most severely, the combination of federal or state dollars and minimum resurfacing requirements appear to have helped turn around urban street preservation. The reinvestment effort should become even more visible in 1985 and 1986 as the contracts awarded under the 1984 FAU provisions are executed in their entirety.

REFERENCES

Eckrose, Roy A., "Measuring Effectiveness of Pavement Preservation Techniques," *Public Works*, July 1979, *110*, 64–68.

Gramlich, Edward R., "State and Local Budgets on the Day after It Rained: Why Is the Surplus So High?," *Brookings Papers on Economic Activity*, 1:1978, 191–216.

______, "An Econometric Examination of the New Federalism," *Brookings Papers on Economic Activity*, 2:1982, 327–66.

Meyers, Harry G., "Displacement Effects of Federal Grants to States for the Primary, Secondary, and Urban Federal Aid Highway Systems," Office of Management and Budget, July 1985, unpublished.

Miller, Edward, "The Economics of Matching Grants: The ABC Highway Program," *National Tax Journal*, July 1984, *27*, 221–29.

Peterson, George E. et al., *Guide to Benchmarks of Urban Capital Condition*, Washington: Urban Institute, 1984

Peterson, George E. and Parham, David W., "Guide to Benchmarks of Urban Capital Condition: 1985 Update," Urban Institute Paper 3302–06, October 1985.

Tye, William B., "The Capital Grant as a Subsidy Device: The Case Study of Urban Mass Transportation," in Transportation Subsidies (part 6), *The Economics of Federal Subsidy Programs*, U.S. Congress, Joint Economic Committee, Washington: USGPO, 1973, 796–826.

Congressional Budget Office, *Efficient Investments in Wastewater Treatment Plants*, Washington: USGPO, June 1985

______, *The Interstate Highway System: Issues and Options*, Washington: USGPO, June 1982

General Accounting Office, *Types of Work Performed Using Resurfacing, Restoration, Rehabilitation, and Reconstruction Federal Highway Funds*, Washington: USGAO, February 1984.

Efficient Pricing and Investment Solutions to Highway Infrastructure Needs

By KENNETH A. SMALL AND CLIFFORD WINSTON*

Premature aging of the interstate system, funding shortfalls, and a well-publicized disaster or two have conspired to bring a crisis atmosphere to discussions of the U.S. highway system. A consensus has developed for making major repairs and establishing financing mechanisms to support them.

This consensus produced the higher truck taxes included in the Surface Transportation Assistance Act of 1982. Although the most weight-sensitive elements were subsequently repealed, Congress directed the Department of Transportation to study the feasibility, efficiency, and fairness of taxes that vary by truck weight and distance traveled. Such efforts may be viewed as attempts to find user charges that bear some relationship to the costs associated with highway wear.

At the same time, highway agencies have sponsored a flurry of activity aimed at finding the best ways to spend rehabilitation funds. This activity has focused on identifying and budgeting for the most urgent needs in a diverse highway system, largely omitting explicit economic optimization considerations. It has also not attempted to consider the economically optimal degree of durability, as measured by road thickness, to be designed into new or rehabilitated roads. This neglect is vital because standard economic theory suggests a close relationship between durability and efficient damage-related user charges.

The purpose of this paper is to discuss the contributions that user charges and investment policy can make toward a long-term strategy for maintaining the load-bearing integrity of the nation's highways. We deal only with costs associated with pavement wear, hence do not consider bridges, congestion, or investments in road capacity.

On the pricing side, we argue that present revenue instruments are inadequate as user charges, and that current levels of heavy-vehicle taxation are too low. On the investment side, we find that past and current design practice leads to serious underinvestment in durability. As a result, the rather high user charges on heavy axles that would now be called for by an efficient pricing rule would be much smaller if investment were optimal. This presents a policy dilemma: is it worth the fight to enact high user charges when ultimately they may not be needed? We argue that this dilemma actually presents the opportunity for a creative policy initiative that meets both short- and long-term needs, and that can be viewed as fair to all sides.

I. Pricing Solutions

An efficient user charge would approximate the social cost of damage that a vehicle inflicts on highways. It is well established that highway damage depends critically upon individual axle weights, varying roughly with their 4th power (U.S. Federal Highway Administration, 1982, p. IV–43). Current attempts to increase taxes on heavy vehicles are unsatisfactory for two reasons: they are based on total vehicle weight rather than axle weights, and they do not vary nearly steeply enough.

In our 1986 paper, we investigate the effects of replacing current mileage-related taxes by an efficient user charge. Based on evidence in U.S. Federal Highway Administration (1982, Appendix E), the user-charge schedule we

*School of Social Sciences, University of California, Irvine, CA 92717, and Economic Studies Program, The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, DC 20036, respectively. This work was supported in part by the Institute of Transportation Studies of the University of California. All views expressed are our own and not necessarily of any institutions with which we are associated.

examine would be higher than current mileage-related taxes for most heavy vehicles, and sharply higher for those bearing heavy loads on few axles. For example, a standard tractor-trailer combination with gross weight of 80,000 pounds would be charged 21 cents per mile, while a three-axle van weighing 55,000 pounds would pay 24 cents per mile.

The efficiency gains are generated by induced behavioral changes, of which we consider two: shipping the same loads in trucks with more axles, and shifting to alternate modes such as rail. (Although buses would also be charged substantially under such a scheme, we ignore them because of their much smaller total number.) We estimate a real resource saving through reduced highway maintenance of some 17 percent of total maintenance expenditures, or about \$2.6 billion annually. This is offset by \$1.4 billion in extra shipping costs and lost consumers' surplus, for a net welfare gain of roughly \$1.2 billion per year. Interestingly, we find that conventional cost-allocation methodology, based on vehicle weight, has misled policy analysts into an undue concern with the heaviest trucks: it is actually the medium-sized trucks (two- to three-axle single-unit vans) from which the greatest gains occur.

The most dramatic effect, however, is a pure transfer of some \$10 billion annually from the trucking industry (or, through forward shifting, from its customers) to the government. The revenue collected from the average five-axle tractor-trailer combination would rise by \$6,000 per year. Political implications are obvious. The new user charge would raise enormous revenues at a time when impending maintenance and rehabilitation needs make new revenue sources an urgent priority; but it would raise strong opposition within the trucking industry.

The magnitude of the transfers inherent in efficient pricing raises two other questions, in light of contentions that roads have been built to inadequate standards. Should the roads have been built so as to better withstand heavy axle weights? And if so, is it fair to penalize the trucking industry for those mistakes? To deal with these questions, we need an optimal investment analysis.

II. Investment Solutions

There are many dimensions over which highway investment can be optimized. These include capacity (for example, number of lanes), durability (for example, pavement thickness), and maintenance and repair strategy. Investment in capacity has been thoroughly studied by economists in connection with congestion pricing (see Winston, 1985, p. 78). Maintenance and repair activities were long neglected, but have recently been subjected to cost-benefit analysis by Jose Gomez-Ibanez and Mary O'Keeffe (1985). Durability, despite a venerable and voluminous engineering literature, has been virtually ignored by economists, except in the case of low-volume roads in underdeveloped countries. This has left design criteria affecting hundreds of billions of dollars to be chosen with no explicit economic optimization model.

Our 1985 paper attempts to rectify that situation. In it, we extend the standard model of pricing and investment developed for congestion to include durability. Optimal durability is calculated by equating the incremental cost of building a thicker pavement to the incremental benefits of reduced expenditures on resurfacing, discounted to the same point in time. We specify empirical relations between capital cost and pavement thickness, and between pavement thickness and pavement life. Pavement life is measured by the number of standardized axle loads (i.e., the passage of a single axle bearing 18,000 pounds) that a pavement can withstand before it must be resurfaced.

This latter relationship was studied as part of a major road test that also provided the 4th-power law, mentioned earlier, relating road damage to axle load. The test was carried out by the American Association of State Highway Officials (AASHO) on test tracks in northern Illinois between 1958 and 1960. Both rigid pavements (portland cement concrete) and flexible pavements (bituminous concrete, commonly known as asphalt) were studied. The results, published in Highway Research Board (1962), have become an important part of the design criteria used by most states (FHWA, 1982, p. IV–42).

The road test had such inherent limitations as a single climate and soil type, a limited range of paving materials, and an inability to test for any independent effects of time and weather. These limitations have engendered voluminous discussion, leading to a reasonable consensus on how to deal with them. By contrast, the primitive nature of the statistical analysis of the road test data has gone virtually unnoticed. Yet by modern standards, the statistical analysis is totally unsatisfactory. Basically, a nonlinear equation describing pavement condition was estimated by dividing the sample into hundreds of separate segments, estimating linear regressions on each, and using the resulting parameter estimates as independent variables in further regressions. Because the linear equations often fit poorly, *ad hoc* adjustments and exclusions were made. Direct comparisons between fitted and actual pavement condition, such as those in W. N. Carey and P. E. Irick (1962) or Canadian Good Roads Association (1962), suggest that this procedure frequently overestimated pavement life. Later evidence, including further observations on some of the test pavements themselves after their incorporation into an actual interstate highway, confirm that the thicker rigid pavements did not last nearly as long as predicted.

One portion of the AASHO equation is of primary interest. It predicts the number of axle loads that can pass over the pavement before its quality reaches a predetermined critical value that mandates resurfacing. This portion was linear in parameters, and we were able to reestimate it using the original data, with the dependent variable equal to actual rather than fitted number of axle loads. We used a limited dependent variable (Tobit) model to take into account those pavement sections that had not yet reached the critical value when the test was discontinued.

Our results show that pavement lifetimes for thick pavements, both rigid and flexible, were substantially overestimated by the AASHO statistical procedures: by nearly a factor of three in the case of a 10″ rigid pavement, the standard currently used in most interstate highways. It is not surprising, then, that many interstate highway sections have in fact deteriorated much more quickly than was expected.

We then used the reestimated equation, along with empirically derived parameters for resurfacing cost and for the incremental cost of thicker construction, to compute optimal pavement thickness. We did this for a real interest rate of 10 percent, and for traffic levels that approximate the 5th, 50th, and 95th percentiles of the actual distribution of traffic levels on six-lane urban interstates in the United States in 1981.

The results indicate that optimal thickness is substantially greater than that provided by current practice. For example, at the median traffic level, optimal thickness for a rigid pavement is 11.5″. This may seem only slightly thicker than the current 10″ standard, but the relationship between pavement life and thickness is so steep than it would last twice as long (26 years compared to 13). Optimal design for a flexible pavement subjected to this traffic would result in a life of 29 years, more than three times the life of a "heavy" flexible pavement as described by U.S. Federal Highway Administration (1983, p. II–10). The divergence becomes even more dramatic at higher traffic levels, and persists over huge variations in key parameters that determine the tradeoff between capital and maintenance (such as a doubling of the real interest rate or of the incremental cost of thicker pavements). In fact, using our basic parameters, the current 10″ standard can be justified only at a real interest rate greater than 20 percent.

Pavements are frequently designed to last 20 years. It appears that such practice suffers from both an inadequate design life, presumably caused by a failure to use explicit optimization, and a discrepancy between design life and actual life, presumably caused by the faulty statistical procedures in the AASHO road test analysis. Another contributing factor is traffic levels that are higher than anticipated.

Because pavement life rises so steeply with pavement thickness (roughly with the 7th power for rigid pavements), these results have dramatic implications for efficient pricing. At optimal pavement thicknesses for the traffic levels we are considering, the marginal cost

of highway wear for heavy trucks is quite low: only 0.8 cents per mile for a single axle weighing 18,000 pounds on a rigid pavement, as compared to 2.2 cents on a 10" and 6.9 cents on an 8" rigid pavement. On flexible pavements classified as "light" in the source mentioned earlier, this marginal cost rises to \$12.75.

Trucking industry representatives are thus correct in claiming that high user charges for heavy trucks would be inappropriate, at least on high-volume roads, if pavements were designed correctly. At the same time, the strong durability economies would still lead to high charges for lightly traveled roads. Thus, efficient pricing would still have an important role in an optimized highway network, but it would not necessarily be one of raising huge revenues. Instead, it would channel the most damaging vehicles onto higher-volume roads, which can economically be built to handle them.

III. Policy Implications

We have argued that given the present design of the nation's highways, efficient user charges would extract large revenues from the trucking industry, and would substantially reduce future repair costs. These features would greatly assist the process of restoring our highway infrastructure to good health. We have also argued that the durability inherent in present design is inadequate, and that optimal durability standards would eliminate the need for such high user charges on high-volume roads. This leads directly to the question raised at the end of Section I: is it sensible or fair to shift the burden for rehabilitating the current inadequately designed system onto the trucking industry?

Unfortunately, we cannot escape the costs of past investment mistakes. Design standards can and should be changed quickly, and the newer standards incorporated into both new highways and rehabilitation projects. But it is obviously impossible to rebuild the entire highway system at once. Meanwhile, the current system will continue to deteriorate, at great social cost, if heavy-vehicle use is not restrained; and revenues for rehabilitation are just as urgently needed regardless of whether rehabilitation could have been postponed through earlier design decisions.

As for fairness, it is fruitless to argue over who "deserves" the blame for highway deterioration. Highways were designed in good faith on the basis of a high quality but imperfect scientific base. The fact remains that the enormous resources expended on those highways are being disproportionately dissipated by heavy trucks.

We therefore advocate the immediate adoption of user charges based on the short-run marginal cost of highway wear. They should vary steeply with axle weights. Insofar as practical, they should also vary by type of road so as to take into account the large differences in vulnerability between thin and thick pavements. For example, a user charge with two rates, one for interstates and one for all other highways, would require only slightly more record keeping than is now used to comply with the weight-distance taxes levied by several states.

Equally important, however, is the adoption of an explicit policy to upgrade the design standards for new or rehabilitated highways. This will reduce future maintenance needs. It also offers a rare opportunity to design a legislative package that meets short-term needs, promotes efficiency, and balances the conflicting claims of fairness in allocating responsibility for the costs of doing so.

To accomplish this, the new user charges must be tied to a toughened design policy so as to fall *automatically* as roads are replaced or upgraded. Because of durability economies, these user charges will not, in the long run, suffice to finance ongoing highway construction and maintenance. Other charges, such as license fees or fuel taxes, will still be needed to cover total highway costs. In this way automobiles, which are far more numerous than trucks but which do negligible damage, will also pay a share of the bill, adding further to the case that equity is being served.

The result would be a temporary but significant increase in the revenues collected from heavy trucks, followed by an ultimate decrease to below present levels. This can be

justified as efficient, fair, and practical. It is efficient because it keeps user charges constantly pegged to marginal cost, and ensures appropriate durability standards for future construction. It is practical because it provides a financing source for urgently needed repairs. It is fair because it assesses the cost of those repairs on the industry most responsible for making them necessary, yet allows that industry to plan for a more satisfactory future. But it will inevitably fail, just as the 1982 federal weight-distance taxes failed, if policymakers treat the new user charges as simply another permanent revenue source.

REFERENCES

Carey, W. N., Jr. and Irick, P. E., "Relationships of AASHO Road Test Pavement Performance to Design and Load Factors," in Highway Research Board, *The AASHO Road Test: Proceedings of a Conference Held May 16–18, 1962, St. Louis, Mo.*, Special Report No. 73, 1962, 198–207.

Gomez-Ibanez, Jose A., and O'Keeffe, Mary M., *The Benefits From Improved Investment Rules: A Case Study of the Interstate Highway System*, Research Report R85-2, John F. Kennedy School of Government, Harvard University, 1985.

Small, Kenneth A. and Winston, Clifford, "Optimal Highway Durability," Working Paper, Institute of Transportation Studies, University of California-Irvine, December 1985.

______ **and** ______, "Welfare Effects of Marginal Cost Taxation of Motor Freight Transportation: A Study of Infrastructure Pricing," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, NBER, Chicago: University of Chicago Press, forthcoming 1986.

Winston, Clifford, "Conceptual Developments in the Economics of Transportation: An Interpretive Survey," *Journal of Economic Literature*, March 1985, 23, 57–94.

Canadian Good Roads Association, *Report of the Observor Committee of the Canadian Good Roads Association on the AASHO Road Test*, 1962.

Highway Research Board, *The AASHO Road Test, Report 5: Pavement Research*, Special Report No. 61E, Washington, D.C., 1962.

U.S. Federal Highway Administration (FHWA), *Final Report on the Federal Highway Cost Allocation Study*, Washington: USGPO, 1982.

______, *Highway Performance Monitoring System Analytical Process*, Vol II: Technical Manual, Washington: USGPO, 1983.

Soviet Growth Slowdown: Duality, Maturity, and Innovation

By STANISLAW GOMULKA*

The purpose of this paper is to offer a consistent interpretation of two apparently contradictory phenomena of Soviet industrial innovation and growth. One is that the average labor productivity growth rate in the period 1928–75 was high, resulting in a significant reduction of the relative labor productivity gap between the United States and the *USSR*. The other phenomenon is that since 1975 this reduction, or the so-called catching-up process, has nearly stopped. The first phenomenon appears paradoxical, since the catching-up process was taking place despite the Soviet system being apparently less conducive to innovation than that of the United States. The timing of the post-1975 slowdown is not self-evident either, since the productivity gap that remains is still large.

1. The Pre-1975 vs. Post-1975 Productivity Slowdown in Soviet Industry

The topic which received much attention in Western literature has been the Soviet growth slowdown in the years 1947–75. However, if one begins analysis in 1928 or 1940, the extraordinary high productivity gains in the years 1947–61 can be interpreted as linked with the significant productivity losses in the years 1941–46. This interpretation rests on the notion that, until 1975, there was in Soviet industry a fairly constant trend rate of innovation and of labor productivity growth, but on this trend there was superimposed a transitory component of technological and (especially) productivity change, negative in the years 1941–46, and positive during the post-1946 reconstruction (see my 1986, book, chs. 7, 8, ff.). The post-1975 slowdown is thus, by this interpretation, of a qualitatively different kind, since it is the trend rate itself which recently appears to have decreased. Moreover, what was extraordinary until 1975 was not so much the productivity slowdown, but the productivity catching up, and therefore this latter phenomenon is what really requires an explanation. I shall suggest an idea which may be useful in providing such an explanation.

The recent innovation slowdown is sometimes interpreted as being caused by errors in the sectoral allocation of investment. Particularly, the transportation sector is suggested to have become a major bottleneck (Herbert Levine, 1983). This argument seems to imply that once the planners note the errors and remove this and other major bottlenecks, the growth rates of labor productivity and output may be lifted and the productivity catching up resumed. At the same time it has been noted by a number of authors, particularly Philip Hanson (1981), that the time periods needed for invention, imitation, plant construction and diffusion are, in the *USSR*, much longer than in the developed West. In addition, there is some evidence of considerable overmanning and underproduction; even when the plants are using identical technology, those in the West employ less workers and produce more output than do those in the *USSR*. In this paper, I will show that these differences in lead times and static efficiency levels are alone capable of explaining much, perhaps all, of the average industrial labor productivity gap between the United States and the *USSR* that is observed at present. Thus, the present relative productivity gap may well simply be the minimum sustainable, or *equilibrium*, gap, given the culture and system-related efficiency environ-

†*Discussants*: Edward F. Denison, The Brookings Institution; Gertrude Schroeder, University of Virginia.

*London School of Economics, Houghton Street, London WC2 2AE, England.

ments in which the innovation, investment, and production activities of the two countries operate.

II. Dualism and the Productivity Catching Up

Dualism is really a catchword to express the notion of an unusually wide variation in labor productivity among production units in less developed economies. For presentation purposes we often imagine these economies as consisting of two distinct sectors, modern and traditional, and thus having a dual structure. The modern sector is presumed to employ a relatively up-to-date technology, usually invented in the world's technology frontier area (*TFA*). This sector typically has a continuous access to foreign technology, through commercial diffusion, its own immitation capability, or other means, access sufficient to keep it relatively close to the world's technology frontier. The traditional sector is based largely on homemade technology; it also tends to employ low-quality labor and low-quality intermediate inputs. At a very low level of development, the size of the modern sector is minor. In my paper (1985), I have developed models to show how the modern sector can serve as an instrument for catching up. Crucial is the investment strategy. At a certain point of development, the sectoral investments are such that employment in the modern sector increases faster than that in the traditional sector. As labor is being transferred to the modern sector, average labor productivity would be increasing from a level prevailing in the traditional sector to a level prevailing in the modern sector, implying a positive growth rate of the average labor productivity even if the sectoral productivities were constant. Suppose the growth rate of the labor productivity in the traditional sector is in fact lower than that in the *TFA*, and the growth rate of labor productivity in the modern sector is the same as that in the *TFA*, the latter rate to be denoted by α^*. As long as the productivity (and technological) gap between the country's modern sector and the *TFA* remains roughly constant, the growth rate of the country's *average* labor productivity would be moving along a "hat-shaped" path (see my book, ch. 3). Simple algebra implies that this rate would be lower than α^* when the bulk of the labor force is still in the traditional sector, but would exeed α^* at a point of time when the size of the modern sector becomes sufficiently large. At that point the catching-up process begins. The speed of that process thus depends on the rate of the country's internal diffusion of modern technologies, which in turn is related to the rate of investment in the modern sector and the supply of sufficiently skilled labor to that sector. The higher that investment rate, the higher also is the rate of internal technological diffusion, and the higher will be the rate of growth of the average labor productivity. In the course of this process there is thus an intimate relation between the so-called extensive and intensive types of growth, the former being an instrument of the latter.

III. The Measurement Problem and the Speed of the Soviet Catching Up, 1940–75

The usual method of estimating the contribution of qualitative changes to the growth of output assumes that the separate contributions to the growth of output of the changes in capital, labor, and technology are mutually independent, a reasonable assumption for small changes in the short run. The method consequently makes no use of the distinction between producible (capital) and nonproducible (labor) inputs, which is however an essential distinction when long-term growth is considered. The point is that technological change, by increasing output in the present period, accounts also for a part of the growth of capital in subsequent periods. The total contribution of technological change (qualitative changes in general) to output growth is therefore higher than the Hicksian productivity residual would indicate. In the typical three-input case—capital, labor, and technology—the total long-term contribution of technological change to the growth of output turns out to be equal to labor productivity growth, provided that, first, the returns to scale are constant and, second, the capital-output ratio remains constant. If these two assumptions hold, average labor produc-

tivity can be used as a measure of the average level of technology. In this case it is unnecessary to estimate the production function in order to assess the contribution of technological (and other qualitative) factors to the growth of output over a long period of time, such as the period 1940–75. This is a most useful implication, given the well-known problems of correct specification and direct estimation of a production function when the data sample is small.

The quality of Soviet industrial data is a well-known problem. However, I find it an inconsistent and misleading procedure to use strongly deflated Western output data and Soviet (or essentially Soviet) capital stock data. This often adopted procedure implies that the Soviet industrial capital-output ratio has been increasing since about 1955 at an annual rate of about 3 percent. Using the Soviet data only, the capital-output ratio index (1960 = 100) stood at 124 in 1940, 97 in 1955, and 120 in 1975. The index has since increased further, but apparently due mainly to lower utilization of the capital stock. These data suggest that for the purposes of long-term analysis the assumption of constant capital-output ratio may be reasonable in the case of Soviet industry. Assuming in addition constant returns to scale and a constant rate of disguised unemployment, I equate the labor productivity proportional changes with the total (direct and indirect) contribution to output growth of all the qualitative changes. This contribution I term the rate of innovation or the rate of technological progress.

The average annual growth rate of U.S. manufacturing output per man-hour in the period 1940–80 was 2.7 percent. According to Soviet official sources, the average annual growth rate of Soviet industrial output per man-hour in the period 1940–75 was 5.6 percent (5.8 percent in the period 1940–60 and 5.3 percent in the period 1960–75). The data imply that the labor productivity catching up was taking place at an average annual rate of 2.9 percent. Taking the index of the U.S.-Soviet productivity ratio as 100 in 1975, the ratio stood at 280 in 1940, 265 in 1950, and 149 in 1960. Using the Greenslade-CIA output data, the ratio would be 188 in 1950 and 122 in 1960. Whether Soviet or Western output data are used, therefore, the rate of catching up has been quite high until 1975. However, the productivity ratio in 1983 was only 98 according to Soviet data and 103 according to the CIA data. It may therefore be assumed that the ratio has been stable since 1975.

IV. Duality Factor in the Soviet Post-1975 Productivity Growth Slowdown

Since 1968, Soviet Narodnoe Khozyaistvo (NK) every 3 or 4 years provides data on the distribution of global output, employment, and capital stock among firms grouped according to their output size category. There are six such categories (in mil roubles, current prices): 0–1, 1–5, 5–10, 10–50, 50–100, and above 100. These data permit the computation of the labor productivity and the capital-output ratio for firms of every output category. I calculated these for the years 1968, 1972, 1975, and 1979 (1979 being the last year for which the data are available). These data give implicitly the distribution of employment with respect to labor productivity. The productivity effect of changes in this distribution is our duality factor.

Labor productivity turns out to increase rapidly with output size, the range being from 1 to 5. The capital-output ratio is about 1.5 times the industrial average for category 1, comprising smallest firms. It declines to about 0.9 times the average for firms in category 4, and increases to about 1.10 for firms in category 6. These data are difficult to interpret unless one assumes that larger-scale firms are, on average, using more productive technology. The data portray a very sizeable shift in employment over time from smaller to larger firms, which may be interpreted as indicating an increasing application of large-scale technologies, which is the diffusion of modern technologies to which I was referring above. Clearly, modern or more productive technology need not be large scale. It would be better for our purposes to have the distribution of employment with respect to labor productivity, whatever the output size of the firm. We do not have these data. However, in Soviet industry, newer may also

be larger. The data we have permit the direct estimation of the labor productivity growth rate due to the shifts in employment from smaller to larger-scale firms. The average annual growth rate of this category was 2.3 percent in 1968–75, and only 1.2 percent in the years 1975–79. The degree of duality in Soviet industry was still high in 1968, there being apparently a major reserve of labor productivity growth until 1975. By 1979, the reserve was evidently largely exhausted, as about 45.7 percent of the labor force and 61.8 percent of the capital assets was already with firms in the two largest output size categories. Incidentally our intercategory shifts in employment did not change the capital-output ratio in 1968–75, but increased it by 1 percent in 1975–79.

V. The Equilibrium Technological Gap

Soviet industrial output per man-hour in 1975 was reported officially to be 55 percent of the corresponding level in the United States (NK, 1983). Independent estimates surveyed by Jaroslav Kux (1976) fall in a range that is uncomfortably wide. Their average would place the ratio at some 45–50 percent for 1975. Assuming the true ratio to be in this range, the labor productivity in Soviet industry in 1975 was achieved in the United States 26 to 30 years earlier.

Hanson reports two surveys of U.K. and German exporters of machine tools and chemical plants: West Germany by Karl Rothlingshofer and Heinrich Vogel; and the U.K. by Malcolm Hill and himself. Their studies produced a wealth of empirical material on comparative lead times and efficiency levels. In particular, the West German executives considered "output levels to average 70–80 percent of normal Western practice, and manning levels to average 120 percent, giving labor productivity levels on the completed plant of about two-thirds of those on similar plant in Western Europe" (p. 200). This evidence is only suggestive. However, if these efficiency rates apply for the whole Soviet industry, then the average labor productivity gap left to be explained by the average technological gap would be reduced to only 11 to 15 years. Technological gaps are bound to vary significantly among product lines, firms, and industries. It is plausible that the gaps are much lower than the overall industrial average in the defense industry, about the average in the investment goods production, and above the average in the consumer goods' industries. These gaps would have to be traced to the various innovation time lags. Consider first the innovations that are imported or imitated. Following the first U.S. application, time elapses before the innovation is noted in the *USSR* (notification lag), before construction of a plan begins (imitation of negotiation lag), before the plant becomes operational (construction lag). There is also a time-lag effect of the slower subsequent diffusion of the innovation in question.

Some of the lags are reported in Hanson. These and other similar data suggest the following lags (in years): notification 1, imitation or negotiation 2, construction 6–8, slower diffusion 3. The total turns out to be 12–14 years, which happens to be within the range of 11 to 15 years of the average technology gap.

For domestically produced innovations, suppose that the research stage begins at the same time in the *USSR* and in the West. Now the stages to be considered are as follows: 1) *R&D*, testing and evaluation, 2) project design, 3) construction, and 4) assimilation and diffusion. According to a Soviet source (Hanson, pp. 79–80), stage 1 takes 87 months, and stage 2, for chemicals, 32 months. We do not have comparable data for Western countries. However, Hanson uses Council of British Industrialists' data for the United Kingdom giving the lead times of some 50–60 months from start of development to start of production for large and innovative projects. Comparable Soviet lead times are reported by Soviet authors to be on average 147 months (Hanson, p. 80). The data imply a 7-year time lag at the point when the stage of assimilation and diffusion begins. Adding 3 years as the minimum difference in the time lag on the latter stage, the total comes to 10 years. (Ivan Semenovich Nayashkov, Chairman of the USSR State Committee for Inventions and Discoveries, gave an estimate worth quoting, although, I

cannot use it directly. In a program on Soviet TV on November 29, 1985, it was suggested to him that "From the practical invention to its introduction it takes over 12 years"; Nayashkov responded: "According to the statistics of the past few years, in our country the time needed on average to bring an invention to life in new equipment is approximately half of what was mentioned by the comrade." (BBC transcript SU/8125/C1/3, 4 Dec 85.)

It should be stressed that this period of 10 years is suggested to be the average time only if the first stage begins simultaneously in the *USSR* and the West. In some cases, Soviet inventors may begin work leading to new technology before a similar program starts in the West. In these cases, the time lag may be smaller than the 10 years, or the *USSR* may even be the leader. However, there must be many other cases where the Soviet *R&D* work begins later than in the West. It seems reasonable to expect that the number of cases in the second category prevails, consequently increasing the average time lag to more than 10 years. It follows that, again, it is possible that the minimum sustainable technological gap for the *USSR* lies in the range 10 to 15 years.

VI. Conclusion

The Soviet and Western data on lead times in invention, innovation, construction, and diffusion appear to imply that the Soviet equilibrium technology gap is 10 to 15 years. If Soviet static inefficiency levels are as high as those reported by Hanson, then the innovation and productivity slowdown in the years since 1975 may be interpreted as the consequence of Soviet industry concluding the catching-up process and reaching its culture- and system-related technological and productivity equilibrium gaps. The state of maturity, in this interpretation, has been reached in which the duality factor no longer plays any significant role.

The other factors—such as bottlenecks in productive infrastructures, a greater share of Siberian investments, greater complexity of the Soviet economy, lower growth of the Soviet *R&D* activity, lower labor morale, and poorer work discipline and effort—may have also contributed to the slowdown. It should be noted that our data on lead times and manning levels relate to the 1960's and early 1970's. We do not need to assume that these leads and levels have since changed for the worse to explain the post-1975 slowdown. However, the quality of these data is such that there is still room for the influence of some or all of these other factors.

REFERENCES

Gomulka, Stanislaw, *Growth, Innovation and Reform in Eastern Europe*, Madison: University of Wisconsin Press, 1986.

______, "Relative Backwardness, Transfer Costs and Appropriate Technology: Models of International Catching-Up," mimeo., 1985.

Hanson, Philip, *Trade and Technology in Soviet-Western Relations*, London: Macmillan, 1981.

Kux, Jaroslav, "A Survey of International Studies of Levels of Labour Productivity in Industry," in F. A. Altman et al., eds., *On Measurement of Factor Productivities: Theoretical Problems and Empirical Results*, Göttingen: Vandenhoeck and Ruprecht, 1976.

Levine, Herbert S., "On the Possible Causes of the Deterioration of Soviet Productivity Growth in the Period 1976–80," in *Soviet Economy in the 1980's: Problems and Prospects*, Joint Economic Committee, Washington: USGPO, 1983.

Soviet Growth Retardation

By PADMA DESAI*

The retardation in Soviet economic growth since 1950 requires explanation. It has also important policy consequences for the Soviet Union, posing choices and raising dilemmas that must be addressed by Mr. Gorbachev as he takes charge of the Soviet economy. I address the former question in depth in the present paper, drawing on the considerable econometric, theoretical and institutional literature that has now emerged on the Soviet economy, touching only briefly on the latter issue for lack of space.[1]

I. Retardation: The Facts

The retardation of Soviet economic growth is a recent, postwar phenomenon. It has been noted by distinguished commentators such as Abram Bergson (1974; 1978) and Herbert Levine (1982). The annual growth rates of *GNP*, estimated as 3-year averages on the basis of the latest Central Intelligence Agency (CIA) data in U.S. Congress, Joint Economic Committee (JEC: 1982), have gone from a high of 7.3 percent in 1955 to 5.4 percent in 1964 and then precipitously and continuously down to 1.9 percent by 1979.

II. Retardation: Causes

In a fundamental sense, the retardation in Soviet growth rates reflects a combination of diminishing returns to capital accumulation and low and even declining rates of technical change (which therefore fail to provide the necessary offset to these diminishing returns). In turn, however, these phenomena must be explained by several factors contributing to plant-level and enterprise-level inefficiency and lack of progressivity as also to economy-wide factors that accentuate these outcomes.

*Professor of Economics, Columbia University, New York, NY 10027.

[1] The complete paper, which examines both the causes and the consequences of the retardation in depth, will be published as chapter 1 in my 1986b book.

A. *Diminishing Returns and Low, Declining Rate of Technical Change*

The Soviet growth experience has traditionally relied heavily on capital accumulation as the source of economic growth, with the "residual" or "technical change" in the sense of Robert Solow (1957) and Edward Denison (1967) playing a limited role in either theory or practice. Soviet planners have thus registered consistently impressive step-ups in rates of saving/investment. These rates have been as high as 25 percent in recent years.

If, then, technical change is low and even declining, and the rate of growth of the workforce is significantly lower than that which would sustain the overall capital-labor ratio in the face of capital accumulation, as is the Soviet case, the decline of growth rates will follow. This growth retardation results from a combination of a shift in the capital-labor ratio that leads to diminishing returns to capital growth, and a low and declining growth rate of Hicks-neutral technical change.

How does one "slice up" the growth retardation explanation between the two contributory factors: low and declining rate of technical change and rapidly diminishing returns to capital accumulation? Two analytical models have been estimated in the considerable econometric literature, one relying on the Cobb-Douglas production function specification and the other on the *CES*. The former implies that greater weight in the explanation is attributed to low and declining technical change; the latter, with a low estimated elasticity of substitution between capital and labor, implies sharper diminishing returns to capital, suggesting that less weight is assigned to technical progressivity as the source of the Soviet malaise.

The econometric studies show that, for data up to 1975, the *CES* production function with a constant rate of technical change, provided the best fit. Thus, Martin Weitzman's (1970) pioneering paper showed this

to be the case for Soviet industry, with an estimated elasticity of substitution of 0.4 and a constant annual rate of technical change of 2 percent. My paper with Ricardo Martin (1983) produced similar results for eight branches of Soviety industry, using both Soviet and CIA output estimates, whereas I (1979a, b) estimated the *CES* formulation with an estimated constant rate of technical change of 1.7 percent and an elasticity of substitution of 0.5 for the Soviety economy as a whole.

But the availability of revised and also later data series up to 1980 has shifted the explanation in favor of the Cobb-Douglas form, with a declining rate of technical change, as originally proposed by Bergson. Thus, I (1985b) used nonlinear estimation techniques, eschewing the competitive relations procedure for obvious reasons as inappropriate in the Soviet context, to show that the rates of technical change in *industry and branches* (with Soviet and CIA data for 1950–80) are declining from year to year and, *when averaged over the period*, are generally low. To quote from my 1985b article: "In particular, the average rate of industrial-output growth in the range of 5.5 to 7.2% between 1960 and 1980 is associated with a negative to poor rate of TFP [Total Factor Productivity] of -1.16 to 0.82% per year. It would seem that the growth of industrial output has been sustained by substantial applications of capital and labor" (p. 21). This holds equally for *economywide* data, again for 1950–80, with the *GNP* data estimated in 1970 rubles by the CIA and reported in JEC (1982, pp. 62–64), the capital stock data in 1973 rubles stated in CIA (1982, p. 3), and the latest labor data in man-hours supplied to me by Stephen Rapawy. With the production function specified as

$$Y_t = Ae^{\lambda_1 t+\lambda_2 t^2}\left(K_t^{\alpha}L_t^{1-\alpha}\right)e^{u_t}$$

where Y, K, and L are the *GNP*, capital, and labor, the estimates of $\ln A$, λ_1, λ_2 and α (with the t-values in parentheses) are: $-0.2006(11.1148)$, $0.0207(2.0047)$, $-0.0004(4.1544)$, and $0.3672(3.0924)$. (Note that the R^2 and Durbin-Watson statistic of the estimated equation are 0.9968 and 1.5316.) When the declining rate of technical change is estimated for each year from these values of λ_1 and λ_2, and averaged over the period, the resulting average rate of technical change is a low 0.89 percent.

B. *Plant-Level and Enterprise-Level Inefficiency*

The foregoing analysis provides the "overall" explanation of the declining growth rate of the Soviet economy: in terms of increasing capital-labor ratios and hence diminishing returns to new investment, and in terms of low rates of innovation which fail to offset the diminishing returns significantly. I now shift focus to what is going on, as it were, behind the face of the clock.

1. *X-Inefficiency.* In the Soviet command economy, enterprises rarely face true competition because there is no threat of entry from either imports or domestic rivals. In this nonpredatory environment, the pressures to put one's best foot forward are missing. The cost curve shifts upward because the sheltered markets obviate the need for cost minimization. The Soviet enterprises, in other words, are plagued by the Leibenstein disease of X-inefficiency: this all-too-natural "goofing-off" effect is probably substantial in the excessively bureaucratized Soviet regime.

2. *Absence of Innovative Activity.* For much the same reason, given the captive buyers and the absent competitors, the inducements to seek and introduce innovations have generally to be weak. Joseph Berliner's (1976) monumental work demonstrates this phenomenon quite effectively.

3. *Inappropriate Incentive System and Decision Criteria.* A great deal of institutional-cum-analytical literature now exists on the fact that Soviet firms and farms operate by criteria different from the profit maximization that would yield economic efficiency within an appropriate economic framework. Among the important contributions to this literature are the work of David Granick (1978), and the analyses by Evsey Domar (1966) of the economics of the Soviet farm and by Weitzman (1976) of the Soviet enterprise under the new incentives. The burden

of this literature is that socialism in practice fails to achieve the economic efficiency that capitalism in theory will and in practice can. And the losses that follow from this particular failing of the Soviet system could be large.

4. *Soft vs. Hard Budget Constraints.* Yet another element of allocational inefficiency at the plant or enterprise level in the Soviet Union has been what Janos Kornai (1982) has recently characterized as soft budget constraints. Enterprises, faced by price-change signals, do not respond by operating according to market economy norms. Rather, they can go to the state for subsidies to obviate the need for adjustment. This dampens the necessary economic response, in terms of production shift, to price changes. It is evidently a source of economic inefficiency.

Kornai is, of course, drawing our attention to the existence in Hungary, the Soviet Union, and other socialist systems of what, in the capitalist system, we might call "political markets."[2] The inefficiency attributable to their existence can be analyzed on the basis of Figure 1 where QQ is the production possibility frontier for goods X and Y. If the commodity price line shifts from P, with production at A, to P_1^H ($= P_1^{S_1} = P_1^{S_2}$), then a fully responsive, hard-constraint system will shift production to B in the usual way. However, suppose now that the economy operates on a soft-constraint system and that producers of Y at A manage to get state subsidies to continue operating at A. Then, assuming (unrealistically) that lobbying for such subsidies is costless in resource use, the economy will shift to the budget line $AP_1^{S_1}$. In this case, the loss from the soft constraint can be measured in the usual Hicksian equivalent-variational fashion as FG in terms of good X. But suppose realistically that the soft-constraint system does use real resources, diverting them from production to subsidy seeking from the bureaucrats as emphasized in the recent theoretical analyses of

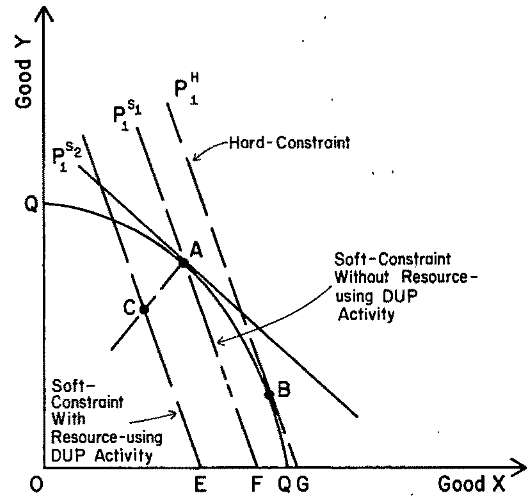


FIGURE 1

such directly unproductive profit-seeking (DUP) activities.[3] In that case, the production vector will shift further in, to say C from A, and the budget line will be yet more impacted and will be $CP_1^{S_2}$. There will then be an added loss of EF. The total loss from operating a soft-constraint socialist system therefore is $EG = FG + EF$, where FG is the loss attributable to reduced price responsiveness, *ceteris paribus*, assuming hypothetically that operating a soft-constraint economy implies no diversion of resources from productive activity, and EF is the additional loss from such resource diversion.

Kornai, who has presumably FG in view, suggests that the loss implied is considerable. No really good estimates are available, either for market economy or socialist economies, on the possible size of EF. Casual empiricism would suggest that the socialist economies do particularly badly on FG whereas pluralistic, lobby-driven market economies do worse on EF. But only systematic probing of this aspect of economic inefficiency will throw light on its importance.

[2] Bergson (1979) also offers an insightful discussion of politics in socialist economies.

[3] See Jagdish Bhagwati (1982) and the substantial theoretical literature which is synthesised and generalized there.

C. *Economywide Factors*

I now turn to an analysis of the factors external to the Soviet enterprise that contribute to diminishing returns to new investments and/or inhibit technical progressivity.

1. *Intersectoral Allocative Inefficiency.* The Soviet Union has a number of institutional constraints, such as on redeployment of capacities and on labor mobility, and imperfections such as differential capital charges that contribute to intersectoral allocative inefficiency. Econometric estimates of this inefficiency have been produced in Judith Thornton (1971) and in my paper with Martin. Developing the distinction between "factor-saving" and "output-augmenting" measures of efficiency, Martin and I devise theoretically defensible estimates of Soviet allocational efficiency. For eight Soviet industrial branches, we conclude that the allocational loss generally ranges from 3 to 10 percent of efficient factor use, with the sensitivity estimates at times climbing up to around 15 to 17 percent. Moreover, these losses rise over time.

2. *Bottlenecks and Shortages.* The Soviet economy is further straight-jacketed by massive rigidities that result in continuous bottlenecks and shortages. While prices are sticky and quantity adjustments are the norm, the all-pervasive phenomenon of queues underlines the failure of the system to adjust quickly to correct the shortages.

The inflexibility of the Soviet economy, arising from restrictions on enterprises in regard to investment and production decisions, is further compounded by the rigidities built into the foreign trade regime. This contrasts sharply with the balance-of-payments practices of the developed market economies in the latter postwar period. Even as the Soviet system has been gradually opened to a higher ratio of foreign trade to *GNP*, the flexibility that would come from enterprises being able to order the needed supplies from abroad has been wholly absent.[4]

As a result, crippling bottlenecks, and associated major losses of output abound in the system. Thus, Gertrude Schroeder (1985, pp. 52–55) writes of shortages of iron ore, coking coal, and hence of steel in 1976–82. She also cites growing bottlenecks in the transportation sector, aggravated by the need to import large quantities of grain, diverting freight capacity from transportation of raw materials.

Similar bottlenecks may have significantly and adversely affected Soviet grain performance. *Weather-adjusted* grain yields, estimated by me (1986a) for 1958–82 show that they have become more variable during 1968–82 than in the earlier years. One of the several explanations is that increased reliance on fertilizer has made Soviet grain yields more vulnerable to interrupted fertilizer deliveries.

The growth of the "parallel," "black," "illegal," or "second" economy in such overly controlled systems provides elements of flexibility as indicated in the recent studies for the Soviet Union.[5] However, the fact that the evasion and avoidance of rules are resource-using *DUP* activities is not customarily taken into account in these analyses, as in Richard Ericson (1983). Once such resource use is allowed for, the final effects of the parallel economy need not necessarily be beneficial.

3. *Marx's New, Compleat Man in Senescence?* It has been suggested, by John Bushnell (1979) and others, that the Soviet economy is plagued by the falling morale of its population. Increased alcoholism is often cited as a surefire index of this. The population, distraught by communism and the stifling of liberties and initiatives, has allegedly been drowning in ever more vodka.

Lassitude, reduced application, and diminished drive can inhibit growth by crippling the efficacy of the labor force. Direct tests of this hypothesis are evidently impossi-

[4] On this issue, see the theoretical analyses of CPE decision making in Franklyn Holzman (1983), and my paper with Bhagwati (1982).

[5] See, for instance, Gregory Grossman (1977). It should be noted that the phenomenon of the parallel economy has been studied by development economists long before the beginning of Sovietological interest in the subject. Thus, see the detailed analysis by Bhagwati and me (1970) for India.

ble without free access by researchers to the population in question. However, insofar as we hypothesise that such inefficiency may show up rather in the form of absenteeism and labor turnover, there seems to be little support for such an assumption. To quote Schroeder: "Soviet sources, in fact, state that rates of labor turnover and absenteeism, while still 'too high', have been *decreasing* since 1970" (p. 67, emphasis added).

4. *Increasing Real Costs*. The last, and perhaps a benign, explanatory factor in the long-run retardation of Soviet economic growth would appear to be the Soviet economy's running into "natural" increases in real costs of producing output at the margin. Imagine going down a mine, extracting successively more intractable layers of coal. Or, recalling Ricardo and Virgin Lands simultaneously, extensive cultivation could require cultivating increasingly less fertile land.

Thus, analyzing and measuring the bias in his factor productivity estimates attributable to these factors, Bergson has noted that "...Soviet final output net of resource costs would have grown by 0.14 of a percentage point...less than if the real price of resources had been constant throughout" (1983, p. 43). In other words, the rate of factor productivity is raised by 0.14 percent to allow for the effects of "natural resource exhaustion." Not surprisingly, this is not a very large effect.

III. Retardation: Consequences

These causes of retardation in Soviet economic growth suggest their own remedies. Evidently, more capital accumulation, through further increase in the already impressive Soviet savings rate, is not the answer. This classic "blood, sweat and tears" route yields increasingly less: as my analysis of the Soviet production functions amply demonstrated.

There are only three possible options open to the Soviet planners: 1) reduce the demands on the pie that is growing ever more slowly: that is, learn to lower expectations and cut the icing from the cake; 2) undertake substantive, extensive and intensive economic reforms that would increase efficiency and continually improve the productivity of resources, thereby raising economic growth; and 3) seek external resources, especially in the form of superior knowhow, that would provide the stream of technical change thereby raising the diminishing growth rate. The first two options offer little promise, making the third critical in Soviet decision making.[6]

[6] In this context, see my article (1985a), and Leonard Silk's (1985) interview with Abram Bergson.

REFERENCES

Bergson, Abram, *Soviet Post-War Economic Development*, Stockholm: Almqvist & Wicksell, 1974.

______, "The Soviet Economic Slowdown," *Challenge*, January-February, 1978, *21*, 22–27.

______, "The Politics of Socialist Efficiency," *The American Economist*, Fall 1979, *24*, 5–11.

______, "Technological Progress," in his and Herbert S. Levine, eds., *The Soviet Economy: Toward the Year 2000*, London: Allen & Unwin, 1983, 34–78.

Berliner, Joseph S., *The Innovation Decision in Soviet Industry*, Cambridge: MIT Press, 1976.

Bhagwati, Jagdish N., "Directly Unproductive, Profit-Seeking (DUP) Activities," *Journal of Political Economy*, October 1982, *90*, 988–1002.

______ **and Desai, Padma,** *India: Planning for Industrialization*, London: Oxford University Press, 1970.

Bushnell, John, "The New Soviet Man Turns Pessimist," *Survey*, Spring 1979, *24*, 1–18.

Denison, Edward F., *Why Growth Rates Differ*, Washington: The Brookings Institution, 1967.

Desai, Padma, (1979a) "The Productivity of Foreign Resource Inflow to the Soviet Economy," *American Economic Review Proceedings*, May 1979, *69*, 70–75.

______, (1979b) "The Rate of Return on Foreign Capital Inflow to the Soviet Economy," in JEC, *Soviet Economy in a Time*

of Change, Vol. 2, Washington: USPGO, 1979, 396–413.

______, (1985a) "Technology Transfer for Mutual Gain," *The New York Times*, February 10, 1985.

______, (1985b) "Total Factor Productivity in Postwar Soviet Industry and Its Branches," *Journal of Comparative Economics*, March 1985, *9*, 1–23.

______, (1986a) *Weather and Soviet Grain Yields*, Washington: International Food Policy Research Institute, 1986.

______, (1986b) *The Soviet Economy: Efficiency, Technical Change and Growth Retardation*, Oxford: Basil Blackwell, 1986.

______ **and Bhagwati, Jagdish N.**, "Three Alternative Concepts of Foreign Exchange Difficulties in Centrally Planned Economies," *Oxford Economic Papers*, November 1982, *31*, 358–368.

______ **and Martin, Ricardo,** "Efficiency Loss from Resource Misallocation in Soviet Industry," *Quarterly Journal of Economics*, September 1983, *97*, 442–56.

Domar, Evsey, "The Soviet Collective Farm as a Producer Cooperative," *American Economic Review*, December 1966, *56*, 734–57.

Ericson, Richard, "On an Allocative Role of the Second Soviet Economy," in Padma Desai, ed., *Marxism, Central Planning and the Soviet Economy*, Cambridge: MIT Press, 1983.

Granick, David, "Soviet Research and Development Implementation in Products: A Comparison with the G.D.R.," in F. Levcik, ed., *International Economics — Comparisons and Interdependences*, New York: Springer-Verlag, 1978.

Grossman, Gregory, "The 'Second Economy' of the USSR," *Problems of Communism*, September-October 1977, *26*, 25–39.

Holzman, Franklyn, "Creditworthiness and Balance of Payments Adjustment Mechanisms of Centrally Planned Economies," in Steven Rosefielde, ed., *Economic Welfare and the Economics of Soviet Socialism*, New York: Cambridge University Press, 1983.

Kornai, Janos, "Adjustment to Price and Quantity Signals in a Socialist Economy," *Economic Appliquee*, September 1982, *25*, 503–24.

Levine, Herbert S., "Possible Causes of Deterioration of Soviet Productivity Growth in the Period 1976–80," in U.S. Congress, JEC, *Soviet Economy in the* 1980'*s*: *Problems and Prospects*, Part 1, Washington: USGPO, 1982, 153–68.

Schroeder, Gertrude, "The Slowdown in Soviet Industry, 1976–1982," *Soviet Economy*, January-March 1985, *1*, 42–74.

Silk, Leonard, "The Pressures in Geneva," *The New York Times*, November 20, 1985.

Solow, Robert M., "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, *39*, 312–20.

Thornton, Judith, "Differential Capital Charges and Resource Allocation in Soviet Industry," *Journal of Political Economy*, June 1971, *79*, 545–61.

Weitzman, Martin, L., "Soviet Postwar Economic Growth and Capital-Labor Substitution," *American Economic Review*, December 1970, *60*, 676–692.

______, "The New Soviet Incentive Model," *Bell Journal of Economics*, Spring 1976, *7*, 251–58.

Central Intelligence Agency, (CIA) *Soviet Statistics on Capital Formation: A Reference Aid*, SOV 82-10093, Washington, D.C., 1982, 3.

U.S. Congress, Joint Economic Committee, (JEC) *USSR: Measures of Economic Growth and Development, 1950–1980*, Washington: USGPO, 1982, 62–64.

Soviet Growth Slowdown: Econometric vs. Direct Evidence

By VLADIMIR KONTOROVICH*

Growth rates of both national income and industrial production in the *USSR* fell from 11–12 percent a year in the early 1950's to 3–4 percent in the early 1980's (*Narodnoe*: 1979, p. 38; 1984, p. 41). Production function studies offer two competing explanations of this trend.

I. Econometric Evidence on the Causes of Slowdown

Studies using the Cobb-Douglas specification blame growth retardation on total factor productivity growth slowdown which set in after 1958. The most comprehensive accounting for the impact of the observable growth determinants (Abram Bergson, 1983) left the decline in total factor productivity growth rate mostly unexplained, and arbitrarily attributed it to technological progress. A number of studies found that the *CES* production function with a constant rate of growth of the residual fits Soviet industrial data better than the Cobb-Douglas production function (for example, Martin Weitzman, 1970). In this case, postwar growth slowdown is explained (at least for part of the period) by decreasing returns on capital under elasticity of substitution below unity and a rapidly increasing capital-labor ratio.

But the estimates of elasticity of substitution and other parameters vary widely across the studies; the implied rate of return on capital in the early 1950's is implausibly high (Bergson, 1979, pp. 117–20; Norman Cameron, 1981, p. 26). Some scholars found elasticity of substitution significantly lower than unity up to the mid-1960's, and close to unity after that (Cameron, p. 36; Ryan Amacher and Darius Conger, 1977, p. 318). Others, in contrast, did not find evidence of a structural break in the sample period (Weitzman, 1983). In some of the latest work, the *CES* production function with less-than-unity elasticity of substitution and a constant rate of growth of the residual was found to fit the data no better, or even worse, than the Cobb-Douglas production function with slowing growth of the residual (Weitzman, 1983; Padma Desai, 1985). Thus, production function analysis of the causes of the growth slowdown is inconclusive.

II. An Alternative Approach

The two hypotheses formulated by production function studies can be evaluated in the light of data that are independent of the time-series used to estimate the production functions. There is no single series that would measure technological progress or diminishing returns; rather I will use a number of indicators that measure important subsets of the phenomena to be analyzed, or that at least strongly correlate with them. Each of the indicators used characterizes the unobserved growth determinant only partially; some are subject to biases. The largest possible number of different indicators is used for each growth determinant so as to cover all its aspects and to check the biases of each individual indicator.

III. Technological Progress

Technological progress here denotes both the first use and the diffusion of new technology. Assuming that the average impact on growth of an innovation does not change over time, and a Cobb-Douglas production function with neutral technological progress, the increment of total factor productivity depends on the number of innovations in a given period. To maintain the rate of pro-

*Command Economies Research, Inc., 2 Sutton Lane, Princeton Junction, NJ 08550. This paper is based on research carried out at the Foundation for Soviet Studies under a contract with the U.S. government. I am indebted to Martin Cherkes, Edward Denison, Wolfram Schrettl, and Gertrude Schroeder for suggestions.

ductivity growth constant, the number of innovations has to increase over time. Data on the number of innovations in the economy and proxies for the impact per innovation will be used, as direct evidence on technological progress.

I use aggregate data from enterprises (basic production units) on innovations implemented during a year in several overlapping categories: planned innovations in industry; inventions (the most radical domestic innovations) and rationalizations (the most numerous, and the least significant, plant-specific innovations) implemented in the economy; number of new types of equipment introduced into and withdrawn from series production; output of equipment (machines, instruments, apparatus) produced for less than 6 years (i.e., embodying new technology). Data on new equipment output are the most reliable, complete, and, arguably, the most significant, since new equipment serves as a conduit for technological progress to all sectors of the economy. The coverage of the other indicators is less complete, but still very broad (hundreds of thousands of planned innovations, millions of rationalizations, and thousands of inventions). Technology transfer from abroad is represented by imports of machinery from the West, as compiled by the Western economists. Taken together, these indicators cover practically the whole universe of technological change (technology transfer from the more developed East European countries being a major omission).

These indicators show a significant slowdown in the growth of the number of innovations over the 1970's. The rate of growth of the number of planned innovations declined from 8 percent in 1971–75 to 4.5 percent in 1976–80, and to 0.9 percent in 1981–82; the number of rationalizations implemented actually declined in 1981–82; the number of inventions implemented declined sharply in 1976–80, though there are some questions about this last piece of data. The number of new industrial products and new types of equipment put into series production grew by 100–200 percent in 1971–75 relative to 1966–70, but by only 3–5 percent in 1976–80 relative to 1971–75 and in 1981–82 relative to 1976–80. The growth of the number of products withdrawn from production decelerated equally quickly over the 1970's. The share of equipment types produced for 5 years or less in the total output of equipment declined from 55 percent in 1967 to 40.6 percent in 1980. Superimposed on the slowdown of equipment output growth in nominal terms in the late 1970's, this produced a marked slowdown in the growth of new equipment output. The average rate of growth of equipment produced for less than 6 years in 1967–78 was 2.5 percent in current prices. This is close to the long-run rate of machinery price inflation, as estimated in the West. Some estimates of the rate of inflation in the late 1970's implicit in the Soviet writings are even higher, entailing an absolute decline in the real output of the newest equipment. Imports of machinery from the West sharply decelerated in the mid-1970's, and according to some estimates, declined absolutely in real terms.

To control for the changes in the average impact of an innovation, I resort to the data on savings from innovations, and on the degree of novelty of new equipment, reported by the enterprises together with the number of innovations. For planned innovations, enterprises report additional profit, labor savings in terms of workers, and savings of current cost and capital. For inventions and rationalizations, only the last is reported. Analysis of planning and reporting procedures, and of organizational arrangements that influence the enterprise behavior, suggests that these data are subject to a number of distortions of unknown magnitude both upward and downward. Growth rates of these indicators will be distorted only if the extent of distortion changes over time (assuming multiplicative distortion). Growth rates will also be influenced by price inflation. The analysis of the economic environment up to 1982 (pressure exerted on the enterprises, priority given to innovations in the plans) shows no reason for the extent of distortion to change. Still, if there were distortions in the rates of growth of savings, these should be only in one direction: exag-

gerating the actual rate, because of the increasing degree of cheating and widening coverage of innovations with savings calculations, and price inflation. This upward bias adds credibility to the decline in the growth rates of savings which was observed from 1970 to 1982.

Indeed, the three measures of savings from planned innovations slowed down more rapidly than the total number of innovations; one of the measures declined in 1976–82. Savings from implemented inventions grew very quickly in 1975–80, which is difficult to reconcile with the sharp decline in the number of inventions; in 1981–82, the growth of savings was within the range of price inflation. The savings from rationalizations did not grow in 1981–82. If even a modest adjustment for inflation is introduced, it becomes clear that savings per innovation did not increase in 1981–82. The share of equipment produced for the first time in the *USSR* in all equipment produced for the first year declined from 47.9 percent in 1973 to 39.5 percent in 1978; that is, new equipment was becoming less radically new.

To summarize, growth in the number of innovations slowed down over the 1970's, and likely came to a halt in 1981–82. The impact of an average innovation did not increase, and may have in fact fallen by the early 1980's. Lack of growth in the number of innovations and their impact means that technological progress depressed the rate of economic growth in 1981–82. It also follows that in the previous decades, technological progress was counteracting growth slowdown. Determining the sign of the impact of technological progress on the rate of growth is as far as we can go, using the data at hand.

The decline or slowdown of each particular indicator can be explained as a purely statistical phenomenon. In most cases, one would not be able to refute such an explanation with the available evidence, but most of these hypothetical explanations appear implausible in the particular circumstances of the Soviet economy in the late 1970's–early 1980's. Also, the fact that all the indicators, compiled independently from one another through different reporting channels, point in the same direction, suggests that they reflect a real phenomenon.

IV. Diminishing Returns on Capital

Diminishing returns on capital are easiest to observe in fixed-proportions production processes. Capital added out of proportion to labor remains idle; the degree to which capital is utilized is readily observable. At the microeconomic level, for a given technology, organization, and skill of workers, capital stock can be defined in terms of workplaces—the number of workers needed for the normal operation of a given production process. Workplaces and workers are inputs to a fixed-proportions production process. The effect on economic growth of adding workplaces in excess of the number of workers available can be taken into account by adjusting the capital stock measure for a declining rate of utilization because of the lack of labor.

Along with the qualitative reports on labor shortage, I use three groups of indicators of the gap between the number of workplaces and the number of workers. The shift coefficient of workers (available for industry as a whole and its sectors) is the ratio of all workers employed during the day to the number of workers in the largest shift. It shows the changes in vacant workplaces in the second and third shifts for discrete production processes. The shift coefficient of machines (available only for the metalworking equipment of machine-building plants) is the ratio of the number of shifts worked by equipment during the day to the number of pieces of equipment installed. It supplements the shift coefficient of workers in that it takes into account workplaces that are not manned even during the largest shift. Ratios of machines to operators for three sectors form the second group of indicators. These ratios are for tractors and harvesters, and trucks in agriculture; bulldozers and excavators in construction; metal-cutting machines in industry and in the whole economy; and presses and forges in industry. For each of these, there exists an optimum ratio of machines to workers, established by the designers of

equipment. When the actual ratio is substantially higher than this engineering optimum, and increasing, this is interpreted as the increase in the share of vacant workplaces. I also utilize direct data on the number of workplaces, mostly provided by I. Malmygin (1984), who pioneered the study of the subject in the *USSR*. While his methodology is not entirely clear, his data are consistent with those from other sources.

The 1950's were the period when the proportion between the workplaces and the workers did not worsen, and may have improved. The ratio of the stocks of two types of metal-working equipment to the number of operators declined from 1954 to 1959. Though the ratio of tractors and combines per operator almost doubled over the 1950's, that of trucks per driver declined. Malmygin characterizes this decade as one of improving factor proportions.

All indicators signal an increasing gap between the number of workplaces and the number of workers in the 1960's. The labor shortage emerged around 1965, according to labor economists. All the ratios of equipment to operators (with one exception) increased by about a third, and that for trucks, by 15 percent. The shift coefficient of workers in industry fell by 9.1 percent from 1959 to 1975, and that of metal-working equipment in machine building fell by 8.3 percent from 1963 to 1971. In his analysis, Malmygin dates the emergence of excess workplaces at 1960 or 1965. Machine-operator ratios for tractors, combines, and for trucks, and shift coefficient of metal-working equipment in machine building stabilize after the early 1970's. However, there are direct indications that the situation in machine building was better than in industry as a whole, where the shift coefficient continued to decline. Also, in the economy as a whole, the number of metal-cutting machine tools per operator increased by 18–25 percent from 1970 to 1980. The share of excess workplaces in the economy increased by 20 percent from 1976 to 1979, according to a Soviet estimate. All this leads to the conclusion that the gap between the number of workplaces and the number of workers continued to increase through the 1970's.

If the excess of workplaces over the number of workers emerged in 1961, and in 1982 amounted to 10 percent of all first-shift workplaces in industry, 9 percent in agriculture, and 16 percent in construction, we can derive the rate of decline in capital utilization in the first shift in 1961–82. To that, we add the rate of decline in capital utilization in the second and third shifts in industry in discrete production processes (representing about half of the total industrial capital stock). This is derived from the changes in the shift coefficient of workers. Making the assumptions about the share of vacant workplaces in the first shift in other sectors (representing a quarter of capital stock), and about the changes of the shift coefficient in nonindustrial sectors, we estimate the decline in the rate of capital utilization due to the lack of workers to average 0.7–0.8 percent per year in 1961–82. This estimate agrees with estimates of Soviet scholars for the 1960's and the 1970's.

Plugging this estimate into Bergson's (1983, p. 37) analysis of the causes of Net Material Product growth slowdown allows us to explain 13–15 percent of the decline in the growth rate of the residual from 1950–60 to 1960–70. The rate of decline in capital utilization in industry is a more reliable figure than that for the whole economy. By plugging it into Weitzman's (1983) Cobb-Douglas production function analysis of industrial growth, we can explain 7 percent of the decline in total factor productivity growth in the 1960's from the 1950's. This does not account for diminishing returns on capital in variable-proportions processes, but the impact of the latter must be smaller: so much capital goes to create new workplaces that there must not be enough capital left for improvement of existing workplaces to result in significant decline of returns.

V. Conclusion

Direct evidence suggests that technological progress has been responsible for factor productivity growth slowdown in the late 1970's–early 1980's. In contrast, identifying the unexplained residual with the impact of technological progress would lead us to blame

the latter for the slowdown through the entire postwar period. Diminishing returns on capital contributed to the slowdown in the latter half of the period, rather than in the earlier half, or throughout the whole period, as indicated by econometric studies. Even in this latter part, diminishing returns are responsible only for a small portion of the slowdown, rather than for all of it. It appears that no single determinant can explain much of the slowdown through the entire 30-year period. Our evidence suggests that there were several such determinants, each responsible for a small part of the slowdown. Moreover, the causes of the slowdown differ from one subperiod to another. More direct data will be needed to draw a reasonably complete list of these causes.

REFERENCES

Amacher, Ryan C. and Conger, Darius J., "Structural Disequilibrium and Growth Retardation in the Soviet Union," *Weltwirtschaftliches Archiv*, August 1977, *113*, 308–21.

Bergson, Abram, "Notes on the Production Function in Soviet Postwar Industrial Growth," *Journal of Comparative Economics*, June 1979, *3*, 116–26.

______, "Technological Progress," in his and Herbert S. Levine, eds., *The Soviet Economy Towards the Year 2000*, London: Allen & Unwin, 1983.

Cameron, Norman E., "Economic Growth in USSR, Hungary and East and West Germany," *Journal of Comparative Economics*, March 1981, *5*, 24–42.

Desai, Padma, "Total Factor Productivity in Postwar Soviet Industry and Its Branches," *Journal of Comparative Economics*, March 1985, *9*, 1–23.

Malmygin, I., "Sistema rabochikh mest: osobennosti razvitiia," *Kommunist*, No. 5, March 1984, 123–25.

Weitzman, Martin L., "Soviet Postwar Economic Growth and Capital-Labor Substitution," *American Economic Review*, September 1970, *60*, 676–92.

______, "Industrial Production," in A. Bergson and H. S. Levine, eds., *The Soviet Economy Towards the Year 2000*, London: Allen & Unwin, 1983.

Narodnoe khoziaistvo SSSR v....g., Moscow: Statistika, 1979 and 1984.

R&D, INNOVATION, AND PUBLIC POLICY

Institutions Supporting Technical Advance in Industry

By RICHARD R. NELSON*

Western economists long have touted our competitive, profit-oriented, market-guided economies as powerful engines of technical progress. Adam Smith thought so. Karl Marx lauded capitalism for this attribute if not for others. Joseph Schumpeter and his followers have hammered the point that this, rather than any tendencies toward Pareto optimality in a static sense, is the hallmark virtue of modern capitalism.

As Sidney Winter and I have argued elsewhere (1982), economists' beliefs that modern capitalism is a fine innovation-generating machine have no intellectual grounding in contemporary neoclassical theory. The twin theorems take technologies as given. The apparent effectiveness of capitalism in this arena must reside in characteristics and mechanisms other than those featured in the microeconomic theory textbooks. In our book, we attempt to analyze these.

It also is true that the capitalist engine is a much more complicated one than many economists seem to think. I presently am engaged in a study of the variety of different institutions supporting technical advance in industry. There is a lot more to the system than, simply, rivalrous for-profit business firms, and patents or alternative mechanisms for appropriating returns. There are, as well, modes of cooperation among firms, and a variety of public institutions dedicated to the generation and spread of technological knowledge. In this paper I report preliminary findings on the roles played by two such institutions; universities and technical societies. But before I do, I should say a bit more generally about the public and private aspects of technology in capitalist economies.

I. The Complex Capitalist Engine

In capitalist economies, technology has two faces—a private and proprietary one, and a public and cooperative one. These at once complement each other, and are at odds.

Schumpeter's analysis of how the capitalist engine works recognizes both sides clearly. He saw the lure and reward for innovation in the quasi rents from a private temporary monopoly. However, in Schumpeter's analysis, the monopoly normally is limited and temporary. Sooner or later, competitors will be able to imitate, or invent around, or develop a better version of, the initial innovation. The fact that an innovation sooner or later goes public has three benefits. Deadweight losses from restriction of use of the particular innovation are kept short run and limited. The innovation can serve as a basis for further innovation by others. And the dangers that an innovator can build a wide and durable monopoly of the industry are kept under control.

Patent law also recognizes these two sides. The inventor gains a limited temporary monopoly in exchange for disclosure that makes the know how public.

From one point of view, the job of institutional design is to get an appropriate balance of the private and public aspects of technology, enough private incentive to spur innovation, and enough publicness to facilitate wide use. But from another point of view, the job is somehow to get the best of both worlds, by establishing and preserving property rights where profit incentives are most effective in stimulating action and where the costs of keeping things proprietary are not high, while

*Yale University, Institution for Social and Policy Studies, New Haven, CT 06520. The research discussed in this paper was supported by the Division of Policy Research and Analysis of the National Science Foundation.

making public those aspects of technology where the advantages of open access are greatest.

II. The Roles of Universities and Technical Societies

Universities and technical societies clearly are important components of the public part of the system supporting technical advance in industry. I report here some preliminary findings as to their roles. The data I present are drawn from a survey, developed and executed by Richard Levin et al. (1984).

A. *Universities*

Universities are a recognized repository of public knowledge. They draw on it in their teaching. They add to it through their research. However, despite the rhetoric in academia about the strong complementarity of research and teaching, the roles of the university as diseminator of public knowledge and creator of new public knowledge are distinct. Academics may be able to teach what new industrial scientists need to know, without having their research be particularly relevant to industry. Basic scientific principles and research techniques may be important, but in many technologies the cutting edge of industrial *R&D* may stand quite separate from what the academics are doing. On the other hand, in some areas of technology, industrial *R&D* appears to be very dependent on academic research.

In our 1984 survey, we asked our respondents to score, on a scale from 1 to 7, the relevance of various *fields* of basic and applied science to technical change in their lines of business. We also asked them to score, on the same scale, the relevance of university *research*. I propose that a high score for a science on the first question signals the importance of university *training* in that field, and a high score on the second relevance of what academic *researchers* are doing.

On the first question, every field of science received a score of 6 or higher from at least a few industries. Four broad fields—chemistry, material science, computer science, and metallurgy—received scores of 6 or higher from over 30 industries (out of 130). The first 3 or these fields received a score of 5 or higher from more than half the industries in the survey. With 5 the cutoff, physics and applied math now make it to the list of sciences scored as relevant at that level or higher by 30 or more industries.

The fact that an industry rated a field of science as highly relevant by no means implies that they rated university research in that field so. While many fields of science are highly relevant to certain industries, and hence university training of scientists in those fields important to technical advance, the relevance of university research to technical advance is more limited or indirect. Thus, while 73 industries rated the relevance of chemistry as a field at 5 or greater, only 18 industries rated university research in chemistry that highly. Forty-seven industries rated the relevance of physics at 5 or greater, but only 3 gave that high a score to university research in physics. It seems likely that industry interest in academic physics departments is concerned mainly with the ability of professors to train future industrial scientists in the basics, and in research techniques. My conversations with a number of *R&D* executives is consistent with this perception.

What fields of university *research* have widespread reported relevance to industry, in the sense that a number of industries accredited university research in that field with a relevance score of 5 or more? Computer science and metallurgy head the list, each with more than 25 industries giving such a score, followed by material science and chemistry, with 19 and 18 industries, respectively.

Biology and the applied biological sciences (medical and agricultural science) warrant special attention. While these fields are deemed relevant by only a narrow range of industries, those industries that scored the fields at 5 or higher almost always rated university research in these fields at 5 or higher, too. Thus those industries whose technologies rest on the basic and applied biological sciences seem to be closely tied to the universities for *research* as well as training.

These findings are consistent with impressions gained from other research of mine concerned with the recent rash of agreements involving corporate support of university research. A large share of these tend to involve the biological sciences. There are, as well, a number of arrangements involving the semiconductor industry. However, these seem directed largely towards supporting university training. Where there is emphasis on research, the apparent industry objective is to make university research more relevant than it has been, as opposed to simply tapping into ongoing research, whereas the latter is the flavor of most of the arrangements in biology.

The findings also are consistent with the answers to another question on the survey. We asked our respondents to score the importance of the contributions of various outside sources to technical advance in their lines of business. The industries giving university research (without specification of field) a score of 5 or higher were almost exclusively those commonly regarded as having technologies based on biology.

B. *Technical Societies*

The principle that technology has a public, as well as a private, dimension is built into the various scientific and technical societies. Generally these societies involve people both from academia and business. They publish journals, hold conferences, and maintain communications networks more generally. In those fields where academic research is important, the societies provide one mechanism through which that research is communicated to an industrial audience and, in turn, academics learn about developments in industry. Technical societies also provide a way for industrial scientists from upstream and downstream industries to meet and exchange information.

We asked our respondents about the importance of the contribution of technical societies to technical advance in their line of buiness. The list of industries that gave technical societies a score of 5 or greater in quite different than the list that rated the importance of university research highly. Most are relatively fragmented industries, and in few of them do the firms themselves undertake much research and development.

TABLE 1—REGRESSION COEFFICIENTS AND *t*-STATISTICS

Dependent Variable	Independent Variables: Univ. Res.	Tech. Soc.
R&D Intensity	.72	−.07
	(2.2)	(−.25)
Contribution of:		
Materials Suppliers	.25	.57
	(1.6)	(4.3)
Equipment Suppliers	.03	.2
	(.23)	(2.5)
Research Equipment Suppliers	.22	.25
	(1.4)	(1.9)

III. How do Universities in Technical Societies Influence Industrial *R&D* and Technical Change?

Few analyses by economists of the determinants of industrial *R&D* and technical change have explicitly considered universities or technical societies as part of the system generating these variables. Here I report tersely preliminary results of a regression analysis that does. In it, the reported contribution of universities, and of technical societies, were included, along with other variables, in equations attempting to explain the *R&D* intensity of firms in a line of business, and the reported contributions of upstream and downstream industries to technical advance in the industry.

The statistics, some of which are reported in Table 1, show the contribution of the university research to be positively and significantly related to the *R&D* intensity of the industry in question, and also positively but less strongly related to the reported contributions of upstream industries. My interpretation is that university research rarely in itself generates new technology; rather it enhances technological opportunities and the productivity of private research and development, in a way that induces firms to spend more both in the industry in question and upstream The contribution of technical societies was strongly correlated with, and

can be interpreted as facilitating, the contributions of upstream firms. It was not significantly related to the *R&D* intensity of firms in the line of business in question.

The hypothesis that the contributions of university research and technical societies work indirectly by enhancing the effectiveness of for-profit *R&D* is supported by statistics from regressions attempting to explain measures of technical change in an industry, which included the research intensity of the industry in question, and the contributions of upstream and downstream industries, as well as the contributions of universities, and technical societies. While all of the former variables have positive and most have significant effects, the contribution of universities and technical societies did not.

REFERENCES

Levin, Richard C., Klevorick, Alvin K., Nelson, Richard R. and Winter, Sidney G., "Survey Research on *R* and *D* Appropriability and Technological Opportunity: Part I," Working Paper, Yale University, July 1984.

Nelson, Richard R. and Winter, Sidney G., *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

The $R\&D$ Tax Credit and Other Technology Policy Issues

By EDWIN MANSFIELD*

As more and more countries, big and small, have adopted $R\&D$ tax credits, the need for additional evidence concerning their effectiveness in increasing $R\&D$ has become obvious and pressing. In this paper, I present the results of what seem to be the first studies of the effectiveness of $R\&D$ tax credits in the United States, Canada, and Sweden, as well as the findings of some other studies bearing on technology policy issues.

I. Effects of Direct $R\&D$ Tax Incentives in the United States, Canada, and Sweden

In 1981, Congress enacted a 25 percent tax credit for $R\&D$ expenditures in excess of the average of a firm's $R\&D$ expenditures in a base period (generally the previous three taxable years). Canada, a pioneer in the use of direct $R\&D$ tax incentives, had both an $R\&D$ investment tax credit and a special research allowance during the early 1980's. The investment tax credit (which was taxable) was 10–25 percent of current and capital spending on $R\&D$, the percentage varying with the size of the firm and the location of its $R\&D$ activity. The special research allowance permitted firms to deduct from their taxable income an amount equal to 50 percent of the increase in operating and capital expenditures for $R\&D$. In Sweden, there was an $R\&D$ tax allowance equaling 5 percent of a firm's $R\&D$ expenditure plus 30 percent of the increase over the previous year.

To obtain information concerning the effects on firms' $R\&D$ expenditures of these direct tax incentives, surveys of the firms were carried out in all three countries. (See my 1985a paper and my paper with Lorne Switzer, 1985.) Stratified random samples (with optimum allocation) of firms were selected. The total sample size was set at 205 firms (110 American, 55 Canadian, and 40 Swedish), since this seemed large enough to obtain the desired precision. As it turned out, the samples included about 30 percent of all company-financed $R\&D$ in the United States and Canada, and about 80 percent of all company-financed $R\&D$ in Sweden. For each firm, an estimate was obtained of the effect, as estimated by the firm's top executives, of the relevant $R\&D$ tax incentive on this firm's $R\&D$ expenditures. The executives who provided these estimates were a random mixture of $R\&D$, financial, tax, and chief executive officers. There was very little problem of nonresponse, and no indication that they felt that their estimates were subject to substantial errors.

The results for the three countries were remarkably similar. Each of these $R\&D$ tax incentives seemed to have increased $R\&D$ expenditures by about 1 (or at most 2) percent. In all cases, the increased $R\&D$ expenditures due to the tax incentive seemed to be substantially (and significantly in a statistical sense) less than the revenue lost by the government. The ratio of the tax-incentive-induced increase in $R\&D$ spending to the foregone government revenue was generally about 0.3 to 0.4. Moreover, in each country, there was substantial evidence that these tax incentives resulted in a considerable redefinition of activities as $R\&D$, particularly in the first few years after the introduction of the tax incentive. Such a redefinition of activities is estimated to have resulted in a total increase in reported $R\&D$ expenditures of about 13 to 14 percent in both Canada and Sweden.

Of course, surveys of this sort must be viewed with considerable caution. Because the Canadian and Swedish tax incentives had been in existence for many years (since 1962 in the case of Canada), it was possible to perform some econometric analyses, based

*Director, Center for Economics and Technology, University of Pennsylvania, Philadelphia, PA 19104. The research on which this paper is based was supported by grants from the National Science Foundation.

on official *R&D* statistics, to estimate the effects of these tax incentives on *R&D* expenditures. While the results are subject to a variety of limitations, they seem to be quite consistent with the survey results. (See my 1985a paper and my paper with Switzer.) Also, other economists seem to have come up with similar results. For example, Charles River Associates (1985), the Congressional Budget Office (1984), and the McGraw-Hill survey of companies' projected spending on *R&D* all have concluded that the tax credit has had relatively little effect on *R&D* spending. Specifically, Charles River Associates has estimated that it has resulted in only a 0.4–0.8 percent increase in such expenditures.

II. Reasons for the Empirical Results

Once one studies in detail the nature of these *R&D* tax incentives, it is easy to see why their effects are so modest. In all three countries, the same factors are at work; to be specific, let us take the U.S. case. For companies that do not raise their *R&D* expenditures over the base amount, the credit is essentially irrelevant. Also, many companies have no income tax liability against which to apply the credit. While they can carry the credit forward to be claimed against future tax liabilities, this lowers its value to the company and reduces its effectiveness. According to Treasury figures, firms in 1981 could use only 59 percent of the *R&D* tax credits, the rest being carried forward.

For companies intending to increase their *R&D* expenditures over the base amount and having an income tax liability against which to apply credit, the credit lowers the after-tax price of an extra dollar's worth of *R&D*. However, for companies intending to increase their *R&D* expenditures each year in the foreseeable future (and this includes most companies because inflation tends to push up nominal expenditures), the tax credit does not reduce the after-tax price of an extra dollar's worth of *R&D* by 25 cents.

Because an extra dollar of *R&D* this year increases the base amount during the next three years, it reduces the credits that will be available then. Assuming that the credit remains in effect for the following three years, the reduction really equals only

$$25\left(1-\frac{1}{3(1+i)}-\frac{1}{3(1+i)^2}-\frac{1}{3(1+i)^3}\right),$$

where i is the interest rate. (See Robert Eisner, et al., 1984, and my 1985a paper.) Thus, if $i=0.15$, the reduction really only equals about 6 cents. Moreover, for firms that do not intend to increase their *R&D* expenditures above the base level or that have no income tax liability against which to apply the credit, the price reduction is far lower than 6 cents.

While our knowledge of the price elasticity of demand for *R&D* is far from adequate, the best available estimates suggest that it is rather low, perhaps about 0.3. If this is so, and if the price reduction due to the tax credit is 6 percent or less, the tax credit would be expected to raise *R&D* by 1.8 percent or less. This compares with the 95 percent confidence interval from my survey for 1983 of 0.6–1.8 percent. Clearly, as indicated above, these survey results are what one would expect, based on the nature of the credit.

III. Proposed Changes in the American Tax Credit

Many economists and policymakers believe that American public policy should promote a greater investment in civilian technology. (See my 1982 book with others.) But the evidence presented above does not indicate that the *R&D* tax credit has had much effect in this regard—nor have the Canadian and Swedish tax incentives. Indeed, in 1984, Sweden eliminated the *R&D* tax allowance, apparently due in part to growing questions concerning the allowance's effectiveness, and Canada changed its system of *R&D* tax incentives in an attempt to improve their performance.

Some economists have argued that, although the effects of the U.S. tax credit on company-financed *R&D* expenditures have

been small, these effects are big enough so that *GNP* is increased by the credit. Others recommend that the government save the $1.5 billion per year that the credit costs in foregone revenue and increase its own *R&D* spending in relevant areas.

My own view is that the *R&D* tax credit should be altered in a variety of ways. For one thing, as the Treasury has argued, the definition of research and development should be tightened. The vagueness of the present definition encourages firms to include as *R&D* many kinds of preproduction expenditures that are not really what informed observers would regard as *R&D*. In addition, the computation of the base amount (currently equal to the average of the company's *R&D* expenditures in the previous three years) should be changed so that there is more incentive for companies to increase *R&D*. In particular, the base amount might be defined as the average amount spent by the company on *R&D* in an initial period times an adjustment factor reflecting industry or national (but not company-specific) changes in *R&D* expenditures since the initial period.

How much effect these changes would have is hard to say, but it seems to me that they are worth trying. According to the study by Charles River Associates, changes in the computation of the base amount (of the sort described above) could result in a substantial increase in the amount of *R&D* induced by the credit. This is because current increases in a firm's *R&D* expenditures would not affect the level of the base amount (and hence reduce credits received) in future years, if these changes were made in the computation of the base amount.

IV. The Leakage of New Industrial Technology

While the *R&D* tax credit has been perhaps the most visible technology policy issue of the past year or two, it has not been the only issue of importance. As is well known, the United States has tried in a variety of ways to stem the outflow to selected countries of defense-related industrial technology. Also, firms such as IBM have been concerned about improper leaks of their technology to other firms. It is no exaggeration to say that the leakage of technology has become of primary importance to both governments and industrial concerns.

Recent empirical studies (see my 1985b article) indicate that both government agencies and firms are likely to encounter substantial difficulties in stemming the technological outflow. According to a random sample of 100 American manufacturing firms, information concerning a firm's development decisions is generally in the hands of at least some of its rivals within about 12 to 18 months, on the average, after the decisions are made. For about one-fifth of the firms, this information leaks out within 6 months, on the average, in the case of product development. Data concerning the detailed nature and operation of a new product developed by a firm are in the hands of at least some if its rivals within about a year, on the average, after the new product is developed. For over one-third of the firms, such information is in their rivals' hands within 6 months. Even in the case of new processes, detailed data generally leak out in less than about 15 months, the major exception being chemical processes, which frequently can be kept a secret for a number of years.

These findings, which seem to be the first systematic data on this basic topic, suggest that American industry is quite porous in this regard. Because employees of one firm exchange information informally with employees of other firms, because engineers, scientists, and managers move from one firm to another, and for a variety of other reasons, technology tends to leak out more quickly than is commonly recognized. Perhaps government agencies and the firms themselves can reduce the leakage rate, but it seems unlikely that they can cut it very sharply.

V. The Role of the Patent System

Having touched on the leakage of technology, it is only a short step to the patent system, which is meant both to facilitate the diffusion of technological information and to protect the incentives for invention and in-

novation. Despite the long history of the patent system, remarkably little is known about its effects. Recent studies have shed new light on the extent to which the rate of development and introduction of inventions would decline in the absence of patent protection. (For example, see my forthcoming article.) According to detailed data obtained from a random sample of 100 firms from 12 manufacturing industries, patent protection was judged to be essential for the development or introduction of one-third or more of the inventions during 1981–1983 in only 2 industries—pharmaceuticals and chemicals. On the other hand, in 7 industries (electrical equipment, office equipment, motor vehicles, instruments, primary metals, rubber, and textiles), patent protection was estimated to be essential for the development and introduction of less than 10 percent of their inventions. Indeed, in office equipment, motor vehicles, rubber, and textiles, the firms were unanimous in reporting that patent protection was not essential for the development or introduction of *any* of their inventions during this period. (For earlier findings, see my article with M. Schwartz and S. Wagner, 1981.)

However, while the patent system seems to have a relatively small effect of this sort in most industries, this does not mean that firms patent only a small percentage of their patentable inventions. On the contrary, they seem to patent about 50 to 80 percent of them, which is testimony to their belief that the prospective benefits from patent protection (including whatever delay is caused potential imitators and the use of patents as bargaining chips) frequently exceed its costs. Even in industries like motor vehicles, where patents are frequently said to be relatively unimportant, about 60 percent of the patentable inventions seem to be patented. Moreover, despite the frequent assertions that firms are making less use of the patent system than in the past, the evidence does not seem to bear this out. This is important because, if the reduction in the patent rate during the 1970's was due to a shift away from patents and towards trade secrets and other forms of protection, policymakers should be aware that this is the case. Based on our results, there is no indication that this is true.

VI. Conclusions

In conclusion, the available evidence seems to indicate that the U.S. *R&D* tax credit, as well as direct *R&D* tax incentives in Canada and Sweden, have increased industrial *R&D* by only about 1 or 2 percent, which amounts to about one-third of the foregone government revenue. To increase the credit's effectiveness, the definition of *R&D* should be tightened, and the computation of the base amount should be changed. With regard to the leakage of technology, the evidence suggests that leaks frequently occur within a year or so, and that it would be difficult to reduce the outflow substantially. Turning to the patent system, it appears that, contrary to popular impression, firms are not making less use of the patent system than in the past. However, only in pharmaceuticals and chemicals is patent protection regarded as essential for the development and introduction of one-third or more of recent inventions.

REFERENCES

Eisner, Robert, Albert, Steven and Sullivan, Martin, "The New Incremental Tax Credit for R&D," *National Tax Journal*, June 1984, *37*, 171–83.

Mansfield, Edwin, (1985a) "Public Policy Toward Industrial Innovation: An International Study of Direct Tax Incentives for R and D," in R. Hayes et al. eds., *The Uneasy Alliance: Managing the Productivity-Technology Dilemma*, Boston: Harvard Business School, 1985.

______, (1985b) "How Rapidly Does New Industrial Technology Leak Out?," *Journal of Industrial Economics*, December 1985, *34*, 217–23.

______, "Patents and Innovation: An Empirical Study," *Management Science*, forthcoming.

______ **and Switzer, Lorne,** "The Effects of R and D Tax Credits and Allowances In Canada," *Research Policy*, 1985, *14*, 97–107.

______ et al., *Technology Transfer, Productivity, and Economic Policy*, New York: W. W. Norton, 1982.

______, **Schwartz, M. and Wagner, S.**, "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, December 1981, *91*, 907–18.

Charles River Associates, *An Assessment of Options for Restructuring the R and D Tax Credit to Reduce Dilution of Its Marginal Incentive*, Boston, 1985.

U.S. Congressional Budget Office, *Federal Support for R and D and Innovation, Washington: USGPO*, 1984.

Longer Patents For Lower Imitation Barriers: The 1984 Drug Act

By HENRY GRABOWSKI AND JOHN VERNON*

On September 24, 1984, President Reagan signed into law the Drug Price Competition and Patent Term Restoration Act of 1984. This law, the first change in United States patent terms since 1861, restores part of the patent life lost during the premarket regulatory process for new pharmaceuticals (and also for medical devices and food additives). A second major provision of the law facilitates the entry of generic competitors after patent expiration.

The adverse impacts on pharmaceutical *R&D* of the 1962 Kefauver Amendents to the Food, Drug, and Cosmetic Act have been well documented (for a survey, see our 1983 book). These regulations have been a significant factor underlying increasing *R&D* costs, longer gestation periods, and shorter patent terms in pharmaceuticals. At the present time, average effective patent life for new pharmaceuticals is approximately half of the statutory life of 17 years. The 1984 Act would increase patent life up to 5 years using a formula approach analyzed in Section II below.

The adverse impacts of regulatory requirements on the entry of generic products in pharmaceuticals have not been well documented. In particular, generic products frequently could not rely on the safety and efficacy evidence submitted by the pioneer firms for post-1962 drug introductions. These data were accorded trade secret status. Consequently, unless the relevant data were publicly available in the scientific literature, an imitator had to duplicate many of the pioneer's tests to gain market approval. Under the new law, a generic drug company need only submit an "Abbreviated New Drug Application" (ANDA). This requires it only to demonstrate that the drug is bioequivalent to the pioneer's product, a relatively low cost experiment.

The Drug Price Competition and Patent Term Restoration Act of 1984 has been termed the most important legislation for the pharmaceutical industry since the 1962 Kefauver Amendments. Essentially, it eliminates duplicative testing and makes entry easy for generic competitors, while at the same time extending patent protection for future new product introductions. In this paper we analyze its likely impacts on competition in the pharmaceutical industry, the incentives for innovation, and general consumer welfare.

I. The Impact of ANDAs on Market Entry

There is evidence to indicate that the requirement that generics duplicate the pioneer's safety and efficacy tests was a significant and growing barrier to entry in recent years. While clinical studies done at medical centers are often published, in-house animal toxicity studies are not. At a minimum, therefore, generic firms had to duplicate these latter studies. It has been estimated in the trade literature that a new drug application for a generic firm potentially involved expenditures of several million dollars and testing periods of 2 years or more.

In order to gain insights into how important this was as an entry barrier, we analyzed 1983 data on the number of drugs among the top 200 pharmaceuticals with expired patents that also had no generic competitors. Antibiotics and pre-1962 introductions were excluded from this analysis because they were governed by separate procedures that did not require duplicate safety and efficacy testing. We found that 34 of the 52 drugs in the sample with expired patents had no generic competition, or 65 percent of

*Department of Economics, Duke University, Durham, NC 27706.

the total. Among the drugs without generic competition were 2 of the top 20 selling products. These 2 drugs had combined sales in 1983 of over $200 million and had patents that expired in 1980 and 1981.

The experience with respect to generic entry for pre-1962 drug introductions and antibiotics was markedly different. We found that over 90 percent of these drugs with expired patents in 1983 had generic competitors. Overall our results indicate that the testing requirement was a significant entry barrier at the time the new legislation was passed.

The Act is particularly timely because there have been a number of complementary developments in recent years operating to encourage increased use of generic drug products. First, all of the state antisubstitution laws prohibiting pharmacists from deviating from physician brand-name prescriptions have now been repealed. Second, third-party payers such as governments, Blue Cross-Blue Shield, and the private health insurers have initiated procedures to limit payment to lower-priced alternatives. Third, the spread of Health Maintenance Organizations has encouraged the use of generic drugs. Finally, consumers have become more conscious of generic drugs as large chain drug stores have aggressively promoted these products (generics generally have higher margins to pharmacists).

The experience of two leading pharmaceuticals, Valium and Inderal, that have experienced generic competition for the first time this year illustrates these trends. These 2 drugs have lost approximately one-quarter of their respective market shares on new prescriptions to generic products selling at price discounts of 20 percent or more. This has occurred within the first 3 months of generic availability. Another leading pharmaceutical, Indocin, has lost approximately half its market share in only its second year of generic competition. These rates of sales losses are far in excess of historical patterns in pharmaceuticals, or what was experienced only a few years ago.

If these numbers are at all representative, the Act has removed a significant entry barrier with enormous financial implications for the pharmaceutical industry and consumers of prescription medicines. In this regard, we calculated that pharmaceuticals with sales of over $2.5 billion in a total market of approximately $13 billion will be subject to market entry via the ANDA procedure during the first year of the new legislation alone. If we take a somewhat longer perspective, 180 of the top 200 pharmaceuticals in 1983 will be subject to generic competition by the end of 1989. This is in very dramatic contrast to the start of the decade when generic competition was largely confined to pre-1962 drug introductions and antibiotics, and "first-mover" advantages were generally strong even for those products experiencing generic competition.

From the perspective of economic welfare, the Act is the source of large potential positive gains along two dimensions. First, it eliminates scientific testing for which there was no valid scientific purpose. Second, it lowers prices significantly to consumers with some elimination of deadweight loss and large transfers from producers to consumers—presumably, a favorable redistribution of income. At the same time, if the Act results in lower market shares and/or lower prices for innovators after patents expire, this could adversely affect the expected returns from *R&D* and lead to lower future drug innovations. The patent term restoration aspects of the 1984 Act are designed to ameliorate this potential situation. Whether the added patent term on future introductions is likely to accomplish this objective is an issue to which we now turn.

II. Patent Extension under the 1984 Act

The Act provides for an extension in effective patent life equal to the sum of the new drug application review time by the FDA plus one-half the clinical testing time, subject to various constraints. These include a maximum extension of 5 years and no extension beyond 14 years of effective patent life. For drugs already in clinical testing, the maximum extension is 2 years. The law also provides a floor of 5 years protection for all new drugs by not permitting any ANDAs during the first 5 years of market life.

In order to obtain some insights into how much extra protection the law will provide, we simulated its effect on the patent life of all the new drugs introduced over the period 1976–81. There were 98 new drug introductions over this period. Their average effective patent life was 8.9 years. If these drugs had been eligible for the full benefits of patent term restoration, the mean effective patent life would have been extended by 2.9 years to 11.8 years. There is considerable variation across the sample with 18 drugs receiving zero extension and 35 drugs receiving the maximum 5-year extension (including the ANDA exclusivity protection for drugs whose patents had already expired by the time of introduction). The annual fluctuation in added patent term varies from 2.2 years for 1978 introductions to 4.2 years for 1980 introductions.

III. The Net Effect on *R&D* Incentives

As discussed above, the Act simultaneously promotes generic competition and restores patent life. In this section we attempt to estimate whether the net effect on incentives to invest in *R&D* is positive or negative. Our approach is to take as our baseline case the cash flow for the average new drug discovered and introduced in the United States in the 1970's. We have described elsewhere (1985) the assumptions and data limitations made in developing this baseline case. We wish to compare the Net Present Value (*NPV*) of the average drug's cash flow for a world without the Act to the *NPV* with the Act in effect. Of course, two key parameters are the number of years of patent life restored, and the percentage loss of net revenues to generics upon patent expiration.

Figure 1 shows the tradeoffs involved. The baseline case has the net revenues life cycle *abcd*. The vertical distance *bc* occurs at time t^* when the patent expires. It represents net revenue loss to generics in a world without the Act, that is, in a world where barriers to entry confronting generic firms are significant. The size of this loss for the average drug is the product of two factors: the probability of entry, and the extent of losses given that entry occurs. We assume an upper bound on this loss to be on the order of 20 percent under the prelegislation regime.

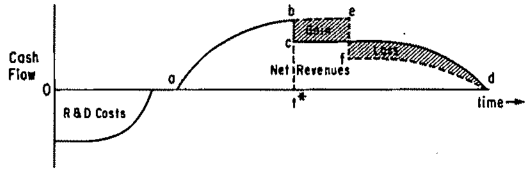


FIGURE 1. EFFECT OF 1984 ACT ON EXPECTED NET REVENUES OF AVERAGE DRUG

In a world with the Act in effect the net revenues life cycle is shown as *aefd* in Figure 1. Of course, the distance *be* represents the patent life restored. The vertical distance *ef* reflects the net revenue loss to generics in a world where generic firms face low entry barriers. Hence, these greater losses to generics as a result of the Act produce the two shaded areas in Figure 1 labled "gain" and "loss." The key issue is then whether the gain or the loss is greater in present value terms.

Given the uncertainty surrounding the two key parameters (the magnitudes of *be* and *ef*), we provide a sensitivity analysis in Table 1. Using 10 percent as the discount rate, the *NPV* for the baseline case is computed and set equal to 100. Three values of patent life restored (1, 3, and 5 years) and three values of net revenue losses to generics (30, 40, and 50 percent) were selected. For the intermediate case (3 years and 40 percent), the *NPV* is shown in Table 1 as only 93 percent of what the *NPV* would have been without the Act in effect. In short, for this choice of parameters, the net effect of the Act is estimated to have a moderate negative impact on *R&D* investment incentives.

Of course, as Table 1 indicates, the net effect of the Act is positive for three cases. For example, with 3 years of restored patent life and net revenue losses of 30 percent, the Act yields a positive effect with the relative *NPV* equal to 102. Although our choice of parameter values is necessarily speculative at this time, it does seem fair to say the tradeoff built into the law is reasonably close —the net effect is not obviously favorable or unfavorable to firms investing in *R&D*.

TABLE 1—NET PRESENT VALUE FOR MEAN DRUG UNDER ALTERNATIVE ASSUMPTIONS ABOUT 1984 LAW'S IMPACT (Net Present Value without Act = 100)

Patent Extension[a] (years)	Net Revenue Loss to Generics[b]		
	30%	40%	50%
5	110	104	98
3	102	93	84
1	91	79	67

[a]Shown as distance *be* in Figure 1.
[b]Shown as distance *ef* in Figure 1.

It should be noted that our analysis has dealt with the average drug. Firms that are optimistically seeking "block buster" drugs are unlikely to be deterred by the Act's possibly negative effect on the *NPV* of the marginal introduction. At the same time, firms that rely heavily on internal funds to fund *R&D* are likely to experience significant financial pressures when the patents on major products expire. The Act will also have very different impacts on firms depending on the diversification of existing drug portfolios, dates of patent expiration, new drugs in the pipeline, etc.

IV. Summary and Conclusions

By eliminating the need for duplicate testing, Drug Price Competition and Patent Term Restoration Act of 1984 should facilitate the rapid entry of generic products. The resulting price reductions to consumers of prescription drugs over the next several years are likely to be substantial. However, the long-run impacts of the Act on innovation are more difficult to assess. Our initial analysis suggests that with an average expected increase in patent life of 3 years, major adverse impacts on the returns to *R&D* are unlikely.

REFERENCES

Grabowski, Henry G. and Vernon, John M., *The Regulation of Pharmaceuticals: Balancing the Benefits and Risks*, Washington: American Enterprise Institute for Public Policy Research, 1983.

______ **and** ______, "Pioneers, Imitators, and Generics: A Model of Schumpeterian Competition in the Pharmaceutical Industry," Department of Economics Discussion Paper, Duke University, 1985.

A New Look at the Patent System

By RICHARD C. LEVIN*

In theory, a patent confers perfect appropriability by granting legal monopoly of an invention for a limited period of time in return for a public disclosure that assures, again in theory, widespread diffusion of social benefits after the patent's expiration. The rationale for this social contract rests on the recognition that technological knowledge has certain attributes of a public good. From this perspective, knowledge, once created, is believed to be freely appropriable by others, and the "free-rider" problem thus limits the incentive to create new knowledge. By conferring property rights that restrict temporarily the wide use of new knowledge, the patent system is supposed to create the incentive to engage in inventive activity and to undertake the costly investment typically required to reduce an invention to practice.

I. Patents in Theory vs. Patents in Practice

This idealized representation characterizes almost all theoretical work concerning the economics of patents. In the recent literature on "patent races" and *R&D* competition, patents are typically represented as providing perfect appropriability, although at least two papers are notable exceptions (Jennifer Reinganum, 1982, and Ignatius Horstmann, et al., 1985). Recent attempts to model licensing behavior, surveyed by Carl Shapiro (1985), similarly treat patents as perfect property rights. On the other hand, theoretical work that takes account of unintended spillovers of knowledge from innovators to rivals (see, for example, Michael Spence, 1984, and my paper with Peter Reiss, 1984) typically pays no explicit attention to the role of patents.

In fact, empirical research, especially that of F. M. Scherer et al. (1959) and C. T. Taylor and Z. A. Silberston (1973), has made it clear that patents rarely confer perfect appropriability. Many patents can be "invented around." Others provide little protection because they would fail to survive a legal challenge to their validity. Still others are unenforceable because it is difficult to prove infringement.

Equally at variance with the theory, unprotected knowledge does not flow freely. Indeed, substantial real resources are often required to imitate an innovation, even one entirely lacking legal protection (Edwin Mansfield et al., 1981). As a consequence, public disclosure of a patent claim does not assure eventual diffusion of the knowledge required to make economic use of an innovation.

The failure of actual patents to conform to the theoretical ideal does not necessarily signal the existence of a policy problem. There is no theoretical presumption that improving appropriability is desirable. Strengthening the patent system may simply reinforce the tendency for patenting to represent a "capture" of property rights, with the associated potential for the dissipation of social benefits through excessive effort to achieve an invention first. Moreover, as a practical matter, powerful incentives to innovate may exist despite the absence of strong patent protection. In the aircraft industry, for example, new products are protected by the inherent difficulty and high cost of reverse engineering complex, multicomponent systems. In the semiconductor industry, where imitation costs are relatively low, returns to new technology are garnered through quick market penetration supported by a steep learning curve. It is by no means obvious that patent protection needs to be strengthened in these two well-studied industries.

*Professor of Economics and Management, Yale University, New Haven, CT 06520. The research discussed in this paper was supported by the Division of Policy Research and Analysis of the National Science Foundation.

II. New Evidence

Until recently, detailed investigation of the effects of the patent system has been confined to a handful of industries. In an effort to develop more comprehensive evidence, my colleagues and I have obtained information from 650 *R&D* executives in 130 different industries. Our survey contained numerous questions concerning the appropriability of returns from *R&D*, as well as questions about the nature of technological opportunity, and the results are described elsewhere (see my 1984 report with others). Here I simply note some conclusions concerning the effectiveness of patents and proceed to discuss some of their implications for public policy.

I focus on one particular set of questions. We asked respondents to rate (on a 7-point Likert scale) the effectiveness of six different means of "capturing and protecting the competitive advantages of new and improved production processes." (We repeated the set of questions for new and improved products.) The listed means of appropriation were patents to prevent duplication, patents to secure royalty income, secrecy, lead time, moving quickly down the learning curve, and sales and service efforts.

We learned that the effectiveness of patents is highly nonuniform across the industries we surveyed. In general, patents were viewed by *R&D* executives as an effective instrument for protecting the competitive advantages of new technology in most chemical industries, including the drug industry, but patents were judged to be relatively ineffective in most other industries.

Consider the 18 industries in which we had 10 or more respondents: pulp and paper, inorganic chemicals, organic chemicals, plastic materials, drugs, cosmetics, petroleum refining, plastic products, steel mill products, pumping equipment, computers, motors and generators, communications equipment, semiconductors, motor vehicle parts, aircraft and parts, measuring devices, and medical instruments. These industries are among the most *R&D*-intensive of the 130 on which we have information, and the average effectiveness of all six means of appropriation was higher in this set of 18 industries than in the full sample. Yet in none of these industries did a majority of respondents rate one of the two patent-related mechanisms as more effective than the most highly rated of the other four means of appropriating returns from new *processes*, although in drugs and petroleum refining a majority regarded process patents as at least the equal of the most effective of the other mechanisms of appropriation. In 5 other industries, one-third to one-half of the respondents thought process patents were no less effective than the most highly rated alternative. Three of these are chemical industries: inorganic chemicals, organic chemicals, and plastic materials.

Patents on new *products* were seen as more effective than process patents in most industries, but only in the drug industry were product patents regarded as strictly more effective than other means of appropriation by a majority of respondents. In 3 other industries—organic chemicals, plastic materials, and steel mill products—a majority of respondents rated patents as no less effective than the best alternative. In several industries producing equipment—pumps, motor vehicle parts, measuring devices, and medical instruments—a significant minority of respondents thought product patents to be at least as effective as other means of appropriation.

In our 1984 study, we analyzed these data in a variety of other ways (Levin et al.), and the conclusions reported here are very robust. Patents were regarded as most effective, absolutely and relatively, in industries with chemical-based technologies. Product patents were seen as moderately effective in a few industries producing relatively uncomplicated mechanical equipment and devices. In most other industries, patents were not viewed as a particularly effective means of appropriation. In addition, we found that patents tend to raise substantially the cost of imitation only in those industries that reported patents to be an effective means of appropriation.

III. Some Preliminary Reflections on Public Policy

These findings raise as many new questions as they settle. For example, if patents are an ineffective means of appropriation in

many industries, why do firms use them? Further study is needed, but one possible answer is that patents are useful for purposes other than establishing property rights. Patents may be used to measure the performance of *R&D* employees, to gain strategic advantage in interfirm negotiations or litigation, or to obtain access to foreign markets where licensing to a host-country firm is a condition of entry.

Suppose it is true that the appropriability of investment in technology is greatly enhanced by patents in chemicals, drugs, and several mechanical engineering industries, but not elsewhere. Some implications for public policy follow directly. In the majority of nonchemical industries, for example, there would be little to gain from lengthening the patent life, since the effect on *R&D* incentives would be negligible. Indeed, even where patents are effective, discounting implies that lengthening the patent life beyond 17 years would not have much impact on incentives unless there is a substantial lag between the grant of a patent and the peak years of its commercial impact (as there is in the drug industry, where the patent life has been lengthened).

Other public policy implications are more subtle and require further study. Consider the treatment of patent exploitation under the antitrust laws. Current law is a woeful tangle of apparently arbitrary and sometimes conflicting doctrines concerning the restrictions that patent holders may impose on licensees. Some practices that are unlikely to have adverse economic impact are illegal per se, although others, potentially more harmful, are subject to a rule of reason. Careful analyis of the efficiency considerations in this area is clearly warranted, and our survey results can inform such an analysis.

To illustrate, note that patent holders seek restraints on licensees to extract more profit from an innovation than could be obtained in the absence of such restraints. In the assessment of any such restraint under the rule of reason, we would wish to discern whether the favorable incentive effect from enhancing the value of patent rights outweighs the anticompetitive effect of the restraint. In such an assessment, general appropriability conditions in the relevant market are an important consideration. If patent protection were inherently strong and imitation costs were high, then restraints on licensees with substantial anticompetitive consequences might be viewed with great disfavor. On the other hand, in markets where patent protection is relatively weak and no other mechanism of appropriation is particularly effective, we might be inclined to take a more permissive posture toward a firm's attempt to extract all it can from a patent.

One more area of public policy deserves mention. During the past year, the Congress and trade policy officials have become increasingly concerned about the possible adverse impact of the intellectual property laws and enforcement policies of some of our trading partners. Considerable attention focuses on infringement by foreign producers of U.S.-owned copyrights on software, books, and audio and video recordings. But there is also substantial support for a policy that would pressure foreign governments to adopt stronger patent laws and to enforce existing laws. Given the absolute and relative inefficacy of patents in many industries, pursuit of these objectives might involve great political cost but generate little benefit to U.S. producers or consumers.

Scrutiny of the specific complaints of U.S. firms, however, reveals that the perceived problems are almost exclusively confined to the treatment of chemical and pharmaceutical patents by foreign governments. Many countries do not permit the patenting of chemical products, although in some of these it is possible to protect a product by patenting the process. Other countries, notably Canada, impose severe restrictions on the exploitation of pharmaceutical patents. Since patents appear to be very important in precisely the industries in which complaints have arisen, efforts to reach some international agreement on appropriate levels of statutory protection and enforcement may be well worth the cost.

REFERENCES

Horstmann, Ignatius, MacDonald, Glenn M. and Slivinski, Alan, "Patents as Information Transfer Mechanisms: To Patent or

(Maybe) Not to Patent," *Journal of Political Economy*, October 1985, *93*, 837–58.

Levin, Richard C., Klevorick, Alvin K., Nelson, Richard R., and Winter, Sidney G., "Survey Research on R and D Appropriability Technological Opportunity: Part I," Working Paper, Yale University, July 1984.

______ **and Reiss, Peter C.,** "Tests of a Schumpeterian Model of *R&D* and Market Structure," in Zvi Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 175–204.

Mansfield, E., Schwartz, M. and Wagner, S., "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, December 1981, *91*, 907–18.

Reinganum, Jennifer F., "A Dynamic Game of *R&D*: Patent Protection and Competitive Behavior," *Econometrica*, May 1982, *50*, 671–88.

Scherer, F.M. et al., *Patents and the Corporation*, 2nd ed., Boston: privately published, 1959.

Shapiro, Carl, "Patent Licensing and *R&D* Rivalry," *American Economic Review Proceedings*, May 1985, *75*, 25–30.

Spence, Michael, "Cost Reduction, Competition, and Industry Performance," *Econometrica*, January 1984, *52*, 101–21.

Taylor, C. T. and Silberston, Z. A., *The Economic Impact of the Patent System: A Study of the British Experience*, Cambridge: Cambridge University Press, 1973.

The Monetary-Fiscal Policy Mix: Implications for the Short Run

By ANDREW F. BRIMMER AND ALLEN SINAI*

In recent years, the policy mix—defined as the contemporaneous joint state of monetary and fiscal policy—has conditioned the patterns of the business cycle, set up numerous imbalances in macroeconomic and microeconomic behavior, and is laying the groundwork for future economic performance. Restrictive monetary and fiscal policies produced back-to-back economic downturns in 1980 and 1981–82. From 1982 to 1985, massive fiscal stimulus against a backdrop of monetary growth targeting by the Federal Reserve comprised a "loose fiscal-tight money" policy mix. Subsequently, an actual and prospective tightening of the federal budget and suspension of monetary growth targeting suggest a shift in the policy mix to a "tight fiscal-easier money" combination.

In this paper, the policy mix of the 1980's first half—in the context of an open economy with flexible exchange rates—is characterized. Some of the important economic and financial effects are identified. Among these are 1) higher nominal and real interest rates than otherwise would have been the case; 2) a strong domestic currency; 3) lower inflation rates; 4) a large and growing trade deficit; 5) an unbalanced composition of economic activity across sectors and industries; and 6) a depressed industrial sector. Some of the changes in economic performance to be expected as the policy mix is shifted in response to the Gramm-Rudman-Hollings balanced-budget statute also are shown.

*Brimmer & Co., Inc. Washington, D.C. 20007; and Shearson Lehman Brothers, Inc, New York, NY 10285 and New York University, respectively.

I. Policy Mix Alternatives, the 1980 to 1985 Episode, and the Analytical Framework

There are four possible policy mix alternatives: loose fiscal-easy money; loose fiscal-tight money; tight fiscal-easy money; and tight fiscal-tight money.

The actual patterns for the fiscal-monetary policy mix since the mid-1950's are illustrated in Figure 1. The ease or tightness of monetary policy is measured by an index constructed as a weighted average of the ratio of free reserves to total reserves, normalized at zero. For fiscal policy, an index based on changes in the full-employment budget deficit relative to nominal *GNP* is used.

During periods of economic slack, a loose fiscal-easy money policy combination has been followed. A loose fiscal-easy money policy is highly stimulative, since it consists of both budget stimulus and rapid growth in bank reserves. When inflationary pressures were dominant, a mixture of tight fiscal-tight monetary policy usually was relied upon.

Loose fiscal-tight money or tight fiscal-easy money policy mixes were never applied in the U.S. economy for any length of time before the 1980's. Between 1981 and 1985, a different policy mix is indicated—one that was loose fiscal-tight money. The genesis of this policy mix lies in the Reaganomics fiscal policies set in 1981 and the new approach to monetary policy—the New Fed Policy (*NFP*)—of October 1979.

Tax reductions proposed by the administration totaled $750 billion over 1981 to 1986. To the current services base of the Carter Administration was added $237.4 billion for defense outlays over the same period.

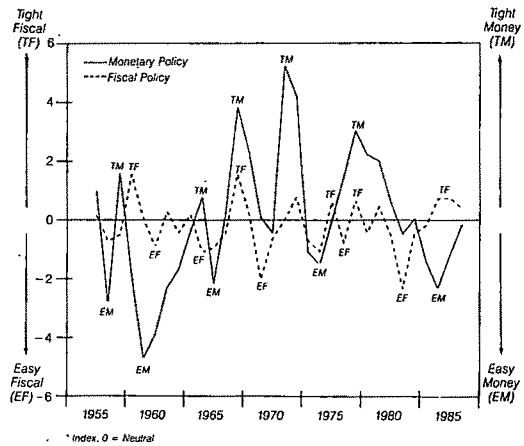


FIGURE 1. POLICY MIX BAROMETER: HISTORY AND FORECAST

Massive reductions in nondefense spending were planned, totaling $585 billion, so that the net difference, tax cuts plus added defense spending less nondefense reductions, was almost $400 billion.

For the first two years of the program, fiscal policy was either restrictive or not stimulative, with a −$2.7 billion net decrement in fiscal year 1981 and only a $12.4 billion increment in fiscal year 1982. From 1983 to 1986, however, some $29.3 billion to $190 billion of fiscal stimulus, measured as the difference between planned injections and subtractions, was indicated.

The essence of the *NFP* was to target monetary growth through bank reserves, freeing interest rates to seek levels consistent with the demand and supply of funds. The deficit stimulus and ensuing heavy Treasury financing, given monetary growth targeting, suggested "permanently" higher nominal and real interest rates as a consequence.

What is to be expected under such a policy mix? Large federal budget deficits suggest increased real economic growth, lower unemployment, more borrowing, and potentially higher inflation. An enlarged *GNP* leads to increased transactions requirements for money, which causes higher nominal and real interest rates. The higher interest rates dampen activity in rate sensitive areas, partially offsetting the fiscal stimulus. In an open economy with flexible exchange rates, stronger growth and higher interest rates strengthen the domestic currency. The inflationary tendencies of the stimulative fiscal policy are mitigated as a result—which further enhances the real returns on financial assets and strengthens the domestic currency, in a kind of "virtuous" cycle.

Because of strong economic growth and a strong currency, imports increase and exports decrease, resulting in a worsening trade deficit. If capital flows impact more than the tradeable goods balance, the domestic currency rises and further worsens the trade deficit. High interest rates also erode the financial positions of households, businesses, and depository institutions, and increase the fragility of the financial system. Debt positions become more onerous, and, if accompanied by falling asset prices, a debt deflation may occur.

So long as huge federal budget deficits remain and other countries do not aggressively stimulate their economies, the process continues until foreign trade weakens so much that domestic economic growth slows. When that happens—the case for the U.S. economy in the second half of 1984 and in 1985—short-term interest rates automatically drop and the value of the domestic currency declines. The eventual result is a lopsided, unbalanced economy, with numerous problems—including chronic federal budget deficits, an overvalued domestic currency, permanent erosion in the relative market shares of the industrial sector, stagnant economic growth, high unemployment, still-too-high inflation, high real interest rates, and a fallout of failures in beleaguered sectors, industries, and financial institutions. Higher inflation is expected and expected real returns on dollar-denominated assets fall. The domestic currency drops in value. Stagflation is an eventual likely possibility. Even a turn toward protectionist policies is a logical consequence.

A tight fiscal-easier monetary policy mix could be expected to have effects opposite to those described above.

II. Policy Mix Impacts—Financial, Macroeconomic, and Sectoral

The federal budget deficits of recent years reached postwar peaks as a result of the fiscal stimulus. On a unified budget basis, the deficit in fiscal year 1984 was \$185.3 billion or 5.2 percent of *GNP*; in fiscal year 1985, it was \$211.9 billion or 5.5 percent of *GNP*. Full-employment budget deficits also indicate considerable stimulus in 1984 and 1985, rising to \$117.1 billion in 1984 from \$77.8 billion the year before, and reaching \$149.7 billion in 1985.

At the root of the persistent federal budget deficits were the large tax cuts adopted in 1981. In that year, budget receipts represented 21.1 percent of *GNP*. By the third quarter of 1985, the fraction was down to 19.7 percent. This lower ratio translates into a net loss of \$54 billion in federal revenues —an amount equal to 28 percent of the \$193 billion NIPA deficit estimated for 1985.

In financing the enormous budget deficits, the federal government has absorbed a growing share of the nation's saving. In 1980, the \$61.2 billion budget deficit represented 15.1 percent of the \$404.8 billion in gross saving in the United States. In the third quarter of 1985, the deficit, at \$214.1 billion (SAAR), was 39.7 percent of the \$506.8 billion of gross saving.

To raise the required funds, the Treasury has become a formidable competitor in the capital markets. In 1980, \$79.2 billion, or 23 percent of the \$334.7 billion raised by all nonfinancial sectors in the United States, was raised. For 1985, the federal share was near 35 percent. If the federal government share had returned to its 1980 position, or 23 percent, the borrowing requirement would have been roughly \$199 billion—or about \$99 billion less than the amount projected.

The impacts of high and rising expected future deficits on long-term interest rates appear in the unusually high levels that have persisted since 1981. The importance of the role played by monetary policy is indicated through a decided break in interest rate behavior before and after the New Fed Policy of October 1979.

The strong federal government competition for funds in the face of monetary growth targeting by the Federal Reserve made interest rates on U.S. government securities considerably more volatile compared with yields on private obligations. Between September 1979 and October 1981, yields on long-term U.S. government and Aaa corporate bonds rose and fell roughly in step. However, since late 1981—when the stimulus of the budget picked up and began to clash against monetary policy—changes in yields on U.S. government bonds greatly exceeded those on corporate obligations.

The strong competition for funds in the U.S. capital markets resulted in a marked climb in U.S. interest rates compared with yields in other countries, providing relatively more favorable returns on U.S. government bonds than the obligations of many foreign governments. In September 1979, long-term U.S. government bonds were yielding 9.26 percent. At the same time, yields in the United States exceeded those in the Netherlands (by 0.2 percentage points), Germany (by 1.2 percentage points), Switzerland (by 5.4 percentage points), and Japan (by 1.0 percentage point). In late November 1985, the margins favoring U.S. government bonds were: Netherlands, 2.78 percentage points; Germany, 3.17 percentage points; Switzerland, 5.01 percentage points; and Japan, 3.36 percentage points.

The combination of a loose fiscal-tight money policy mix and higher interest rates produced a very sharp appreciation in the dollar. At the peak of its appreciation in February 1985, the U.S. dollar was 62 percent above the March 1973 benchmark and 92 percent above the trough in 1980. By September 18—just prior to the joint decision of the United States, France, Japan, Germany, and the U.K. to intervene in the foreign exchange markets—the dollar had depreciated by 12.4 percent. By mid-December 1985, a further dollar depreciation of 11.8 percent had occurred.

With strong growth in the U.S. economy in 1983 and 1984 and continuing rises in the dollar, it has been no surprise to see massive declines in the U.S. merchandise trade bal-

ance and in real net exports over the last few years.

In 1980, imports of merchandise represented 9.3 percent of the U.S. *GNP*, and exports equaled 8.4 percent. The foreign trade gap was −\$24.1 billion, or 0.9 percent of *GNP*. In the third quarter of 1985, exports had shrunk to 5.4 percent of *GNP*, and imports were 8.9 percent. The corresponding trade deficit had risen to −\$138.1 billion—equal to 3.5 percent of *GNP*.

Another impact of the policy mix has been much lower inflation, with dollar appreciation having a surprisingly large impact on U.S. inflation during 1981 to 1985. Considerable slack in the U.S. and world economies permitted the rise in the dollar to drive inflation lower by anywhere from 3 to 6 percentage points.

The sectoral performance of the U.S. economy also reflects the policy mix. Substantial rises in consumption and business fixed investment were registered, an apparent consequence of the tax cuts for households and businesses. Between 1982:4 and 1985:3, business capital spending rose 34.8 percent, well in excess of the average 22.9 percent rise over the same period of expansion in four previous episodes. Consumption outlays were up 14.2 percent, somewhat greater than the average experience in most other expansions.

III. Policy Mix and the U.S. Industrial Sector

The large appreciation in the dollar from July 1980 to February 1985 made U.S. exports extraordinarily expensive and created difficulty in competing abroad. Foreign products valued in other currencies have been relatively cheap. The differentials sparked an enormous expansion of imports into the United States and greatly diminished the role of the U.S. industrial sector in the world economy.

The pervasive rise in imports has had an adverse impact on employment in the United States. From January 1980 through November 1985, total employment rose by 8 million persons. However, this net gain was the product of a 9.4 million increase in employment in service-producing industries and a reduction of 1.4 million jobs in goods-producing industries. Within the latter, manufacturing jobs declined by 1.5 million, jobs in mining decreased by 42,000, and those in construction rose by 174,000.

The reductions in manufacturing employment were especially large in those industries subject to intense foreign competition. In percentage terms some of the decreases were: basic steel, 45.6 percent; textile mill products, 20.4 percent; fabricated metal products, 13.1 percent; and nonelectrical machinery, 15.7 percent. In contrast, employment in printing and publishing (mainly protected from imports) rose by 14.6 percent.

A more detailed view of the adverse impact of the trade deficit on the U.S. economy is provided by an examination of changes in import market shares, production, and employment for 72 Department of Commerce industry groups over the first half of the 1980's.

In 1979, imports accounted for an average of 11.6 percent of the domestic market of the 72 industry groups. The import market share ranged from a low of 0.9 percent (automotive stampings) to a high of 45.8 percent (rubber and plastic footwear). In 18 of the industry groups (one-quarter of the total), imports held under 5 percent of the domestic market, and in only 6 industries did the import share exceed 30 percent. By 1985, imports satisfied an average 16.4 percent of domestic demand across the 72 industry groups, ranging from a low of 1.7 percent (also automotive stampings) to a high of 53.4 percent (radio and TV receiving sets). In only 9 industries was the import share under 5 percent. It exceeded 30 percent in ten industries. By 1985, there were 33 industry groups where the import share was 15 percent or above—compared with only 12 in 1979.

The share of imports in the U.S. domestic market rose in 61 of the 72 industry groups between 1979 and 1985. Among the 61, the actual level of domestic production in 33 (54 percent) was below that achieved in 1979. The level of output was higher in 27 industry groups (44 percent) and unchanged in one case. In the 10 industry groups where the

import market share declined, 8 saw a rise in output while only 2 experienced a decrease in production.

The response of employment to changes in import market shares was quite different. Employment fell in 44 (72 percent) of the 61 industry groups in which the market share of imports increased. In 15 cases (25 percent), employment increased despite the climb in imports. In the 10 industries where the degree of import penetration decreased, 5 experienced a rise in employment, and 5 saw a reduction.

IV. The 1985 Twist in the Policy Mix and Future Implications

Two doses of budget tightening were effected during 1985, and the changed policy mix will modify the profile of economic performance. The first was from a compromise between the administration and Senate Republican leaders in early May when substantial reductions in spending were agreed upon. The second occurred as a result of the deliberations leading to and resulting in the adoption of the Gramm-Rudman-Hollings (GRH) deficit reduction legislation.

The GRH legislation established a process by which the deficits are to be gradually phased out by fiscal year 1991, either through voluntary measures on spending and taxes agreed upon by the Congress and administration, or involuntarily through automatic spending cuts. The administration and Congress can employ any means to reduce the deficits to the targeted ceilings, but, if agreement is not reached, across-the-board spending cuts in eligible categories of nondefense and defense will be made to achieve the target.

During 1985, monetary policy also shifted. The central bank made clear that monetary growth targeting would no longer be followed, given a break in the relationship between monetary growth, especially for $M1$, the rate of inflation, and real economic growth. With monetary velocity declining and off its trend path, the central bank officially decided to relegate monetary policy to a lesser position. The economy, inflation, and the dollar were elevated in the list of considerations for policy.

TABLE 1—EFFECTS[a] OF GRAMM-RUDMAN-HOLLINGS: (1986–90 Average)

Category	No Fed Ease	$M1$ Restored	*GNP* Restored
Fed. Gov't Spending (Bils. $'s)			
Ex Ante	63	−63	−63
Ex Post	−89	−99	−110
Real *GNP* (% Chg.)	−1.0	−0.5	0.1
Nominal *GNP* (% Chg.)	−1.4	−0.9	0.0
Final Demand (Bils. '72 $'s)			
Constant	−1.2	4.1	10.8
Bus. Fixed Invest.	−2.5	−0.1	3.1
Resid. Invest.	1.6	2.7	3.8
Net Exports	5.0	3.9	2.5
Housing Starts (Mils. of Units)	.070	.112	.152
Aaa Corp. Bonds (Basis Pts).	−151	−168	−189
Fed. Funds Rate (Basis Pts.)	−145	−221	−294
Deficit (Bils. $'s)	71	89	109

Source: Computer simulations with the Shearson Lehman Model of the U.S. economy. Econometric model simulations can only produce approximate outcomes—the central tendency of a distribution of possibilities given the change in policy and structure of the economy as assumed and estimated in the model, provided that the policy shock is within the range of historical experience.

[a]Changes from baseline.

With prospective significant reductions in budget deficits seemingly set for coming years and a change in the approach to monetary policy by the Federal Reserve, the policy mix was twisted sharply toward a tighter budget and easier money.

Major differences in the pattern for the unified budget deficits, Treasury financing, and economic performance can be the result of the GRH legislation. Under GRH, the unified budget deficits could be near $180 billion in fiscal year 1986, about $150 billion in fiscal year 1987, and $120 billion or so in fiscal year 1988. Total Treasury financing would decline from the $197.3 billion in fiscal year 1985 to $120.6 billion in fiscal year 1988, compared with $170.1 billion for fiscal year 1988 before GRH. The structural budget deficits under GRH would decline sharply in fiscal years 1986, 1987, and 1988, indicating significant fiscal restraint. Before the GRH legislation, the structural budget deficits also

were expected to decline, but by much less. As a percentage of *GNP*, the full-employment budget deficit could drop to 1.0 percent by fiscal year 1988 instead of the 2.5 percent indicated prior to the legislation.

Some economic and financial market effects of GRH were simulated with the 300-equation Shearson Lehman Model of the U.S. economy (see Table 1). In the first case, it was assumed that the Federal Reserve does not provide any compensating easing. Under these circumstances, the tighter budget has a negative impact on real output, consumption, business fixed investment, and federal spending. However, interest rates are lower, the housing sector improves, and net exports are higher (col. 1).

In the second case (*M*1 restored), the Federal Reserve returns *M*1 to the pre-GRH path. The results include a smaller reduction in nominal and real *GNP* and in business fixed investment. The interest rate declines are larger, and the expansion of homebuilding activity is stronger. More of a cutback occurs in federal spending, and the gain in real net exports is somewhat weaker (col. 2).

Finally, in the third case (*GNP* restored), it was assumed that the Federal Reserve acts to return *GNP* to the pre-GRH baseline. This produces the most favorable outcome. In response, interest rates declined the most, and rate sensitive sectors (particularly housing and investment) show the most improvement. Consumer expenditures rise more strongly, and federal spending decreases relatively more from large savings of interest payments. On the other hand, the expansion of real net exports is less (col. 3).

V. Concluding Comments

This paper has analyzed the monetary-fiscal policy mix in the early 1980's and its implications for financial markets, the economy, and the U.S. industrial sector in the short and intermediate term. For much of the early 1980's, the policy mix could be characterized as a loose fiscal-tight money combination.

There are many messages to be drawn here. Perhaps the main one is that the policy mix, especially in an open economy with flexible exchange rates, must be monitored, analyzed, and understood as much as the individual state of monetary or fiscal policy itself, in order to grasp the likely patterns of behavior in the financial markets and the economy. The policy mix is an important ingredient in the business cycle. A similar message would apply to the rest of the world as well, not really considered here, with like patterns of behavior to be expected under similar configurations of the policy mix as have been analyzed for the U.S. economy.

The Monetary-Fiscal Mix and Long-Run Growth in an Open Economy

By FREDERICK C. RIBE AND WILLIAM J. BEEMAN*

The widespread concern about large government deficits is based upon the crowding-out of productive capital, and eventual reduction in living standards, that such a fiscal outlook portends. As a number of authors have noted, however, there has been little quantitative content to these arguments, so it has been hard to judge how serious the consumption implications of the deficit are.

A recent paper by Edward Gramlich (1984) provides such analysis using a long-run growth model. (Long-run models are appropriate because the main issue involves capital accumulation; such analysis may have little empirical value, but it is useful in a normative context.) Gramlich's most striking result is that a period of 20 years or more might pass before an increase in the deficit that is assumed to represent increased government-provided consumption produces enough of an erosion in the growth of capital and output to cause an offsetting decline in private consumption. Gramlich, however, omits any consideration of international capital flows and of variations in monetary policy, features that seem central to such quantitative analysis.

This paper presents results from an open-economy (two-sector) growth model with explicit fiscal and monetary policies in both the U.S. and rest-of-world sectors. The results imply that the drop in consumption caused by sustained large government deficits in one sector may be significantly smaller in an open economy with active monetary policy than was suggested in Gramlich's paper. Other disturbing consequences emerge, however.[1]

I. The Model

There are two sectors, in each of which one good is produced using a Cobb-Douglas technology identical to that in the other sector, involving the technical-progress-augmented "effective" labor force. Dynamics are in discrete time. The growth rate of the effective labor force and all other behavioral parameters are assumed to be the same in both sectors. The difference in scale between the sectors is accounted for by assuming that the effective labor force in the rest of the world is R times that in the United States.

Saving in each sector is by the life cycle model, making it a linear function of the local stock of wealth and of (the labor share of) local output. The real interest rate in each sector is linearly related (by the rental-price expression) to the marginal product of capital located there.

The flow of saving (in each sector) results in a stock of financial wealth that is allocated among that sector's outside money, interest-bearing claims on the local government and capital stock (which are assumed to be perfect substitutes), and interest-bearing securities issued by the other sector, which are not necessarily perfect substitutes for local securities. The share of wealth allocated to each asset is assumed to be a linear function of the model's two nominal yields: that on local securities, which is the local real interest rate plus an expected inflation rate that is discussed below, and a similar yield on the other sector's securities adjusted for expected

*Congressional Budget Office, Washington D.C. 20515. The views expressed here are our own and do not necessarily reflect positions of the Congressional Budget Office. We thank Frank deLeeuw, Jacob Dreyer, Victoria Farrell, and George Perry for helpful discussions, and Stacy Miller and Jeffrey Steger for expert programming.

[1] The results presented below are quantitative counterparts to material in Peter Diamond (1965). For a more detailed account, see our paper (1985).

exchange rate changes, which are also discussed below. Thus, using the subscript d to refer to the local sector and f to refer to the other sector, we have

$$(1)\quad M_d/W_d = a_m + b_m(i_d) + c_m(i_f')$$

$$(2)\quad (K_{dd}+D_d)/W_d = a_k + b_k(i_d) + c_k(i_f')$$

$$(3)\quad K_{df}/W_d = a_f + b_f(i_d) + c_f(i_f'),$$

where M_d is the stock of outside money in sector d, W_d is its stock of wealth, and $(K_{dd}+D_d)$ is the sum of local wealthholders' interest-bearing claims on the local capital stock and the local government. K_{df} is local wealthholders' interest-bearing claims on the other sector, which are assumed without loss of generality to be claims on its capital stock. i_d is the local nominal interest rate, and i_f' is the exchange-adjusted nominal rate in the other sector. The usual Brainard-Tobin adding-up constraints apply to the coefficients.

The real capital stock in a given sector is the sum of the holdings of local capital by wealthholders in the local and in the other sector. Holdings by the latter are determined in part by equation (3) as it applies to the other sector, and in part by the real exchange rate, since portfolio holdings are valued in terms of the currency of the holder, rather than that of the issuer. Thus,

$$(4)\quad k_d = k_{dd} + k_{fd}(ep_f/p_d)\ \text{(scale factor)}.$$

Here k_d is the real capital stock in sector d, k_{dd} is local holdings of this sector's capital as determined in equation (2), k_{fd} is the other sector's holdings of local capital as determined in equation (3) as it applies to the other sector, e is the nominal exchange rate expressed in terms of units of local currency per unit of the other sector's currency, p_f is the price level in the other sector, and p_d is the local price level. The indication "scale factor" refers to the fact that the differential scales of the two economies are relevant here.

The price levels. Equation (1) is inverted and used to determine the price level as a function of the (exogenous) nominal money stock.

The exchange rate is determined by the balance of flow demands for foreign exchange arising from exchanges of real goods and services and of financial claims between sectors. Real imports of goods into a given sector are assumed to be unit elastic with respect to both output and the real exchange rate.

The balance-of-payments identity on the basis of which the exchange rate is determined is

$$(5)\quad (K_{df}-K_{df-1}) - K_{df-1}i_f + Im_d = \left[(K_{fd}-K_{fd-1}) - K_{fd-1}i_d + Im_f\right]e.$$

Here, K_{df} is sector d's claims against the other sector and K_{fd} is the obverse, both from versions of equation (3) but expressed here in aggregate nominal units of the home currency of the holder of the claim. i_f is the other sector's nominal interest rate not adjusted for expected exchange rate appreciation. Im_d and Im_f are real imports of goods by sectors d and f, respectively, in aggregate nominal units of the demanding country's currency.

Fiscal and monetary policies. Fiscal policy in each sector determines that sector's primary budget deficit (the overall deficit less outlays for interest) and monetary policy determines the growth rate of the money stock. These variables in turn affect the evolution of that sector's stock of interest-bearing government debt (and therefore the degree of crowding out) through the familiar difference equation

$$(6)\quad d_d = pd_d + d_{d-1}\frac{(1+i_d)}{1+G_d} - m_{d-1}\frac{mg_d}{1+G_d}.$$

Here, pd_d is the primary deficit, d_d is the stock of government debt, m_d is the real money stock, G_d is the current nominal growth rate of the economy, and mg_d is the growth rate of the money stock.

Expectations. In the simulations reported below, expectations of inflation and exchange rate depreciation, which affect the nominal interest rates, are the model's own predictions during later periods, representing the assumption of perfect foresight. Solutions are computed using a three-stage iterative procedure due to Ray Fair and John Taylor (1983).

II. Parameterizing the Model

Most behavioral parameters were given values equal to or consistent with the corresponding parameters in Gramlich.[2] The most sensitive parameters, however, are the b_i's and c_i's that appear in equations (1)–(3), reflecting the degree of substitutability between foreign and domestic financial assets, and these have no counterparts in Gramlich. In the computations we used values for these parameters based on empirical work with the capital asset pricing model (*CAPM*) implying a high, though still finite, degree of long-run substitutability.[3] We also did simulations assuming no intersectoral mobility of capital at all.

As initial conditions, a steady state was constructed in which the levels of the variables in each sector were assumed equal to each other at levels roughly reflecting conditions in the United States at the outset of the 1980's (Table 1).

TABLE 1—VALUES OF MODEL VARIABLES IN INITIAL STEADY STATE[a]

Variable	Value
Real Output[b]	1.000
Real Capital Stock[b]	3.668
Real Government Debt[b]	0.296
Real Wealth[b]	4.300
Real Interest Rate	0.010
Nominal Interest Rate	0.040
Inflation Rate	0.030
Exchange Rate	1.430
Depreciation Rate[c]	0.000
Price Level	1.000
Primary Budget Deficit	0.022
Money Growth Rate	0.056

[a]Corresponding variables equal or roughly equal across sectors.
[b]Per effective worker.
[c]Exchange rate.

III. Numerical Results

Three 40-period simulations were conducted in which the model was shocked out of this steady state by a permanent increase in the U.S. primary budget deficit of 1 percent of *GNP*—roughly the size of the increase that occurred in the early 1980's in the United States. The rise in the deficit was assumed to finance government-provided consumption. In successive simulations it was assumed that the deficit increase was accompanied by: no intersectoral capital flows ($c_m = c_k = a_f = b_f = c_f = 0$); intersectoral capital flows (portfolio parameter values estimated using the capital asset pricing model); and, finally, intersectoral capital flows and an increase of 0.1 percentage point in the U.S. money-growth rate.

In each solution the ratio of U.S. government interest-bearing debt to wealth roughly doubled, causing crowding out of holdings of real domestic capital from U.S. financial portfolios. When no intersectoral capital flows were assumed, the consequences closely resembled those reported by Gramlich: 18 years passed before real consumption per capita fell below its initial level. This period was more than doubled (to 38 years) when

[2]The capital share in output, accordingly, is 0.26. The marginal propensity to consume out wealth is .155 and that out of income is 0.3, values that are consistent with Gramlich's 0.8 average propensity to consume. The elasticity of demand for real imports with respect to income and the real exchange rate are both taken to be unity for partially technical reasons. The growth rate of the augmented labor force is 2.5 percent. The real depreciation rate in both sectors is 6 percent, and the multiplicative wedge relating the real interest rate to the marginal product of capital is 1.01. The ratio of the scale of the rest-of-world sector to the United States is 1.5, reflecting recent OECD figures on output in member countries. These figures are also quite close to those assumed or implied in James Tobin's paper presented in this session.

[3]Using annual IMF data, we developed a weighted-average, exchange-adjusted yield on government securities in non-U.S. OECD countries, and computed the variance-covariance matrix of this yield with the corresponding U.S. yield. The *CAPM* implies that the portfolio coefficents are derivable from this matrix (Benjamin Friedman and V. Vance Roley, 1979).

intersectoral capital flows were assumed in the second simulation, however, and it was extended further with the expansionary U.S. monetary policy that was assumed in the third simulation. Since consumers were assumed to treat the accumulating government debt as wealth, the gradual decline in labor income as crowding-out proceeds was the only major factor causing consumption to fall.

Faster U.S. money growth increased consumption by reducing the degree of crowding out through monetization and by temporarily increasing the U.S. nominal interest rate (through higher inflation) relative to the exchange-adjusted nominal rate in the other sector, attracting more foreign capital. The real exchange rate appreciated gradually in all the simulations.[4]

Gross domestic product per effective worker in the United States declined by 1.3 percent over 40 years in the second simulation relative to the level it would otherwise have assumed. This is about one-third as much as in the closed-economy run. The same variable in the foreign sector declined by one-half percent as U.S. crowding-out spread to it. However, gross *national* product in the second simulation declined more sharply in the U.S. (1.8 percent), and less sharply in the foreign sector (0.2 percent) as foreign claims to U.S. output increased. These developments may provide cause for policy concern.[5]

[4]In this model, unlike others (for example, William Branson, 1985), an increase in the current account deficit is required to approach a new steady state. This accounts for the real appreciation.

[5]In another simulation we increased the budget deficit by the same amount in both sectors, assuming intersectoral capital flows were present. The relatively severe crowding-out experienced by the United States in the closed-economy simulation above occurred in *both* sectors, with only small capital flows.

REFERENCES

Branson, William H., "Causes of Appreciation and Volatility of the Dollar," paper presented to the Jackson Hole Conference, Federal Reserve Bank of Kansas City, August 1985.

Diamond, Peter, "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, *55*, 1126–50.

Fair, Ray C. and Taylor, John B., "Solution and Maximum Likelihood Estimation of Dynamic Nonlinear Rational Expectations Models," *Econometrica*, July 1983, *51*, 1169–85.

Friedman, Benjamin, and Roley, V. Vance, "A Note on the Derivation of Linear Homogeneous Asset Demand Functions," NBER Working Paper 345, May 1979.

Gramlich, Edward, "How Bad are the Large Deficits?," in Gregory B. Mills and John L. Palmer, eds., *Federal Budget Policy in the 1980's*, Washington: Urban Institute, 1984, 43–78.

Ribe, Frederick, and Beeman, William, "Fiscal Policy Effects in an Open-Economy Growth Model," unpublished, Congressional Budget Office, 1985.

The Monetary-Fiscal Mix: Long-Run Implications

By JAMES TOBIN*

Since 1981 the United States has faced, for the first time in history, the prospect that the federal debt would grow faster than the national product indefinitely. Economists have been prominent among the Cassandras deploring runaway public debt, but they have not been very specific about its hazards to the health of the nation. The usual story is "crowding out." The citizens' savings are limited. The more that the federal government borrows, the less are available for capital investments, the sources of productivity advances on which the living standards of the future depend. This is an unexciting story of slowdown in growth, and we are usually imprecise about magnitudes and speeds, and especially about whether, how, and when government borrowing leads to a catastrophic crisis.

The purpose of this paper is to present explicitly and precisely the crowding-out story, in a way that exposes the roles of the parameters of fiscal and monetary policies and the macroeconomic structure. The model is first presented algebraically, and then illustrated numerically by simulations assuming arbitrary, but, it is hoped, plausible parameter values.

For many sets of parameter values, these simulations do end in catastrophes, which can be precisely described and dated. I want to be clear at the outset that these are illustrative exercises, warnings—not predictions. They are, it is true, motivated by recent trends in the United States. But my prediction is that tax and/or expenditure policies will sooner or later be changed enough to stabilize the debt/*GNP* ratio. Indeed, those changes appear to be on the way already.

*Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520. I thank Daphne Butler and Willem Thorbecke for valuable help in computations.

I. The Model: General Structure

Since I am here concerned with long-run trends, the structure of the model is borrowed from neoclassical growth theory. Full employment of an exogenously growing labor force is assumed. Gross output is produced by labor and fixed capital; it is divided between consumption and investment by the saving decisions of household members who are both workers and capitalists. The production function allows for variable proportions of the two inputs. The marginal productivity of capital determines the short-term interest rate.

Government debt *is* private wealth in this mode. That is why it crowds out the alternative store of value, productive fixed capital. Ricardo-Barro equivalence effects, whatever their general validity, are not appropriate here. No one expects deficit-reducing tax increases or spending cuts when none have been put in place and the government denies or ignores their necessity. The message of this exercise is that they do need to be put in place so that people have credible reason to expect them.

Transactions with the rest of the world are not modeled in this exercise. In practice, as shown by recent experience, crowding out of net foreign investments (current account surpluses) mitigates the impact of budget deficits on domestic capital formation. But borrowing abroad cannot in general spare the economy the consequences of allowing government debt to absorb ever increasing shares of private saving. It could do so only if foreigners were willing to lend to us indefinitely at real interest rates below our economy's trend rate of real growth.

There is nothing new in studying long-run crowding out by use of neoclassical growth models. That is commonly done by comparative static analysis of steady states. Reduction in the fraction of national product saved and invested, resulting in the case at hand

from governmental dissaving, moves the economy slowly from one steady-state path to another. The second path has a smaller capital-labor intensity than the first, therefore lower real wages and lower per capita consumption. Neither the transitional dynamics nor the differences between equilibrium paths are very dramatic. Students to whom this scenario of crowding out is exhibited yawn and wonder what all the shouting is about.

What is different in my simulations is attention to the possibility that, for quite realistic values of parameters, no steady states exist, or that the only one close to the initial conditions is an unstable equilibrium. In these cases an unstable vicious circle can lead fairly quickly to a dramatic crisis.

The key departure from the usual projections of deficits and debt growth and their effects is to make interest rates endogenous. Interest costs contribute to deficits and the growth of public debt. Increases in rates are the mechanism by which government borrowing squeezes capital investment.

Monetary policy enters the model via the fraction of public debt monetized by the central bank. Here I assume that the Federal Reserve holds the inflation rate constant and determines the degree of monetization accordingly. A higher inflation target slows down the pace of crowding out, because greater "seignorage" lowers the interest cost on total debt. In this full-employment model, there are no direct monetary effects on capital formation, which is governed wholly by saving. Given the fiscal parameters, the only way the central bank can alter policy is to change its inflation target.

There are two parameters of fiscal policy. The main fiscal parameter is the ratio of the *primary deficit*—the deficit exclusive of after-tax interest outlays—to *GNP*. Given this parameter, changes in the income tax rate make a difference in two familiar ways. Private saving depends on after-tax income, and both saving and money demand may depend on after-tax interest rates.

II. The Model in Detail

The nonfederal sector of the economy holds at all times a stock of wealth equal to a multiple μ of the *GNP*. Life cycle theory suggests that nonhuman wealth is a multiple, w, of after-tax labor income, which here is a constant fraction $1-\alpha$ of *GNP*, where α is the elasticity of gross output with respect to capital in a standard two-factor Cobb-Douglas constant-returns-of-scale production function. Then, writing the federal proportional income tax rate as τ and letting $u=1-\tau$, we have

$$(1) \qquad \mu_0 = wu(1-\alpha).$$

Wealth demand may also be related to the after-tax interest rate uR. Here this relation is assumed linear; a nonnegative coefficient β represents the response of savings to interest.

$$(2) \quad \mu = \mu_0 + \beta u(R - R_0) = \mu_0' + \beta uR.$$

(R_0 is introduced simply for calibration of the initial conditions of the simulations. Other variables similarly subscripted below play the same role.)

The wealth-*GNP* ratio μ is composed of the capital-*GNP* ratio k and the government debt/*GNP* ratio d:

$$(3) \qquad k + d = \mu.$$

Debt takes two forms: nonmonetary, costing the government an after-tax real rate of interest r; and monetary, costing the government zero nominal interest, thus a real rate the negative of the inflation rate π. The stock of base money is a fraction h of *GNP*. The demand for base money, relative to *GNP*, depends on the nominal interest rate $r+\pi$. For a given inflation target π the central bank sets h to meet the demand:

$$(4) \qquad h = h_0 - \gamma(r - r_0) - (\pi - \pi_0) = h_0' - \gamma(r + \pi).$$

The nonnegative coefficient γ is higher the more sensitive are demands for base money to nominal after-tax interest rates.

The fundamental dynamic equation for d is

$$(5) \qquad \dot{d} = x + d(r - g_y) - h(r + \pi).$$

Here x is the primary deficit in ratio to GNP. The growth rate of real GNP is g_y. Equation (5) says that the debt grows by the primary deficit if there is no outstanding debt at all (first term); that given an initial debt, it grows further by its net interest cost to the Treasury but declines relative to GNP by the economy's growth rate (second term); and that it declines by the amount of seignorage (third term).

In a steady state, real output can grow at its natural rate g. The gross marginal productivity of capital is α/k, and δ is the constant rate of capital depreciation:

$$R=\alpha/k-\delta. \tag{6}$$

The government's net interest cost of borrowing is lower than R for two reasons. One is that its creditors return part of their interest receipts in taxes on the interest—indeed on the nominal interest, a fact that saves the Treasury $\tau\pi$. The second is that the government can borrow with a credit-risk discount v from the after tax return to capital equity. Thus

$$r=uR-v-\tau\pi. \tag{7}$$

The actual growth of GNP is

$$g_y=g+(\alpha/(1-\alpha))(\dot{k}/k). \tag{8}$$

Combining (5) and (3) yields a differential equation in k:

$$\dot{k}=\dot{\mu}-x-(\mu-k)(r-g_y)+h(r+\pi). \tag{9}$$

This is the fundamental dynamic equation of the model. The strategy for its solution is to express all the variables in (9) in terms of $\dot{k}$ and k and of the policy parameters (x, u, π) and structural parameters $(w, \alpha, \beta, \gamma, \delta, g, v)$. The other equations above enable this to be done.

Tedious algebra leads to the differential equation:

$$\dot{k}=Q(k)/V(k), \tag{10}$$

where $Q(k)$ is a cubic polynomial and $V(k)$ is a quadratic, with coefficients that depend on the parameters. If β and γ are both zero, the degrees of the two polynomials are reduced by one. This is a convenient reference case.

The more important of these two conditions is the assumed zero value of β. This implies that crowding out is unrelieved by any increase in saving induced by rise in interest rates. A positive value of γ, on the other hand, tends to make matters slightly worse than in the reference case. As interest rates rise and the demand for money falls, the central bank has to monetize less debt in order to meet its inflation target. In any case, for realistic values of h, seignorage is quantitatively small.

III. The Solution

Figure 1 depicts Q, V, and $\dot{k}=Q/V$ for the reference case, all as functions of k. Since $d=\mu-k$ and $\dot{d}=\dot{\mu}-\dot{k}$, it also depicts $\dot{d}$ as a function of d. The quadratic Q has a positive intercept. Its roots are its intersections, if any, with the horizontal axis. In Figure 1 there are none. The denominator V is a negatively sloped line, which crosses the k axis at $\alpha\mu$. The vertical line at that point separates two quite different behaviors of $(\dot{k}, k)$. To its left, $\dot{k}$ is always positive, rising asymptotically to the dividing line. This part of the solution has no practical interest. The relevant region is to the right of the dividing line, where Q/V will have roots if and only if Q does. If these exist, they are steady-state values of k and thus also of d. The higher of the two is stable, the smaller unstable. If, as in Figure 1, Q has no roots, $\dot{k}$ is always negative in the relevant region; there is no steady state. Clearly, whether or not Q has real roots depends on the parameters. The general case, with β and γ nonzero, is qualitatively like the reference case but more complicated.

The economy is in trouble if there are no positive real roots of Q and thus no steady states. From whatever initial condition in the right region, k will steadily dwindle, at an ever increasing rate. The same trouble occurs even if roots and steady states exist, if the initial values of k and d are to the left of the lower, unstable root. The question is, what happens as k declines along the curve Q/V, according to which $\dot{k}$ goes to minus infinity.

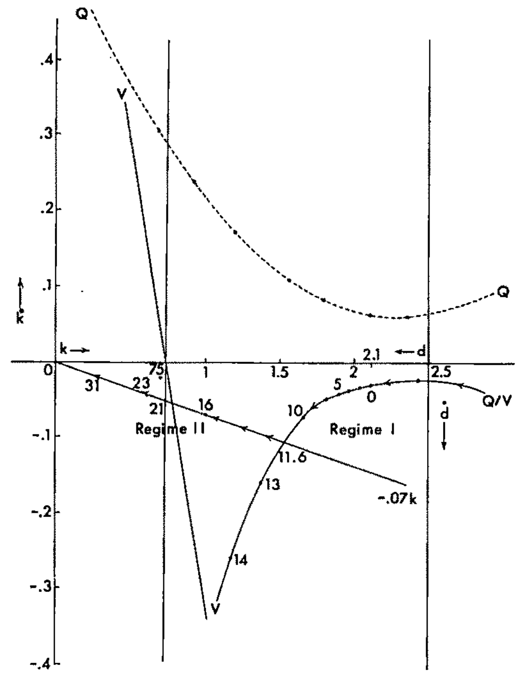


FIGURE 1. SIMULATION OF THE MODEL (REFERENCE CASE)

Note: Numbers refer to numbers of years from "present." Regime I years 0–11.6, Crowding Out. Regime II years 11.6–, Capital Consumption.

Gross investment cannot be less than zero; the capital stock cannot decline faster than the depreciation rate δ. Accordingly, the capital-output ratio k cannot decline faster than

$$\dot{k} = -(\delta + g)(1-\alpha)k. \tag{11}$$

Equation (11), a line graphed in Figure 1, replaces (9) when Q/V is smaller than this number. The intersection of the two functions is the point of switching from regime I to regime II.

Although the rate of capital consumption is limited, the growth of debt continues. In regime II, $k + d$ exceeds μ. What gives? Assuming the structural and policy parameters remain the same, the natural adjustment is the valuation of the capital stock. In regime I, the value of q was implictly 1. But in regime II, when no gross investment is taking place, $q<1$. Indeed such valuation is the neoclassical signal to agents that investment is an uneconomic use of output. The wealth constraint becomes

$$qk + d = \mu, \tag{12}$$

$$\dot{d} = -\dot{q}k - q\dot{k} + \dot{\mu}. \tag{13}$$

The equation for the interest rate r is also different. Holders of public debt compare its return not with uR, but with $uR/q + \dot{q}/q$, what they can earn on equity. Thus (6) becomes

$$r = u\alpha/qk - u\delta/q + u\dot{q}/q - (v + \pi\tau). \tag{14}$$

The dynamics of regime II can be found by substituting (11)–(14), into (5), the same strategy used above for regime I. The result is an expression giving $\dot{q}$ as a function of q and k:

$$\dot{q} = Q(q, k)/V(q, k), \tag{15}$$

where, given k, Q, and V are, respectively, quadratic and linear in q. In regime II the growth of debt "crowds out" q, pushing it down to zero, to make room for d within the wealth-income ratio.

IV. Simulation Results

The parameter values used in the simulations are as follows:

Policy Parameters: x (primary deficit/GNP) = 0.03; τ (tax rate) = 0.20; π (inflation target) = 0.04.

Structural Parameters: α (capital share of GNP) = 0.30; δ (depreciation rate) = 0.07; g (natural growth rate) = 0.03; μ (private wealth/GNP) = 2.50; h (base money/GNP) = 0.05; β (savings/interest coefficient) = 0; γ (money demand/interest coefficient) = 0; v (interest premium) = 0.01.

Initial Conditions: $d(0)$ (debt/GNP) = 0.40; $k(0)$ (capital/GNP) = 2.10; $R(0)$ (real interest rate (pre-tax) = 0.073.

Table 1 exhibits selected results and Figure 1 shows the simulation graphically. The initial condition for $(\dot{k}, k)$ is the point marked 0 on the curve Q/V. From that point k declines, d increases, as shown on

TABLE 1—SELECTED SIMULATION RESULTS

	Regime Change	$q=0$
T (Year)	11.6	23
$k(T)$ (Capital/GNP)	1.5	0.7
$d(T)$ (Debt/GNP)	1.0	1.8
$Y(T)/Y'(T)$ (GNP/Natural Growth GNP)	0.87	0.63
$C(T)/C'(T)$ (Consumption/ Natural Growth C)	1.0	0.76

the curve for subsequent dates 5, 10, 11.6. At time 11.6, there is a regime switch. The regime II path is just a line from that point to the origin.

Figure 2 shows what happens in regime II. The upper left panel shows the decline in q and qk as d continues to increase, until, when q is zero at time 23, the debt absorbs all private wealth. The upper right panel tracks the amount by which debt plus capital stock valued at par exceeds the demand for wealth. The lower panel shows $\dot{q}$ over the same period.

Thus the debacle is the cessation of gross investment, followed by a decline in the stock market until the surviving capital stock is valueless.

Table 1 includes the simulated values of capital stock/GNP, debt/GNP, GNP itself, and national consumption (including government purchases) for the two critical dates, for regime shift and for $q=0$. The GNP and consumption are measured relative to what they would have been had they grown steadily at the economy's natural rate of growth g. Note that before the regime change, the shortfalls of those variables are trivial; indeed consumption has not suffered, because the decline in national saving has offset the decline in capital intensity. As I remarked above, the visible penalties of gradual crowding out are undramatic. The crunch comes in regime II, when losses of output and consumption become severe.

How do variations in the policy parameters affect the outcomes? The key fiscal parameter is x, the ratio of the primary deficit to GNP. With the other parameters the same as in the reference case, x must be below 0.00375 to make the system stable.

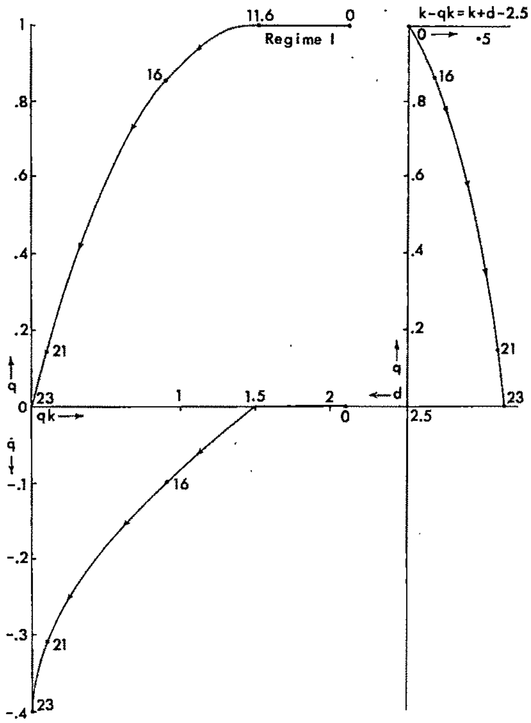


FIGURE 2. SIMULATION OF REGIME II (REFERENCE CASE)

Note: Upper left: Decline of q; Lower panel: Path of $\dot{q}$; Upper right: Excess supply of capital.

But with x at that value, the steady-state values of k and d are, respectively, 2.37 and 0.13. The initial conditions (2.1, 0.4) are on the wrong side of that equilibrium. Hence k dwindles, but so slowly that the economy is still well inside regime I after 50 years. To have a steady-state equilibrium k of 2.1 or lower, x must be smaller than -0.001. In effect, a balanced primary budget is stable at or near the initial conditions. That is true in the reference case because the net real interest rate on federal debt, allowing for monetization, is very close to the natural growth rate. This choice of parameters and initial conditions was not accidental; it appeared to me that the U.S. debt/GNP ratio could be stabilized if the primary budget were balanced.

The inflation target makes a difference too. But in the reference case, with $x=0.03$, it takes a 12 percent inflation to obtain a stable

solution. Likewise, lowering the target from the reference value of 4 to 1 percent shortens the life of regime I by only one year.

I turn to the structural parameters. The most important one is β, the responsiveness of private demand for wealth to the real interest rate. In current macroeconomic theory, the interest elasticities of saving and wealth demand are key parameters, and there is a lively debate about their empirical magnitudes. In the reference case, perfect inelasticity is assumed. At the other extreme, perfect elasticity at the initial interest rate, clearly there would be no crowding-out problem at all.

An informative summary measure of the effects of varying a parameter is the duration of regime I, in years. In Figure 3, this measure is related to the value of β. To calibrate β, note that a value of 10 corresponds to an elasticity of about 1/4. The figure assumes the reference case values of other parameters. Given so high an x, strict stability is not possible with any finite β. But positive values of β do slow down the crowding-out dynamics and prolong regime I. Values of 15 or higher make its duration longer than 25 years. The interest elasticity of money demand is less important. Introducing a γ of 0.25, which corresponds at initial values to an interest elasticity of money demand of about 1/4, turns out to make very little difference.

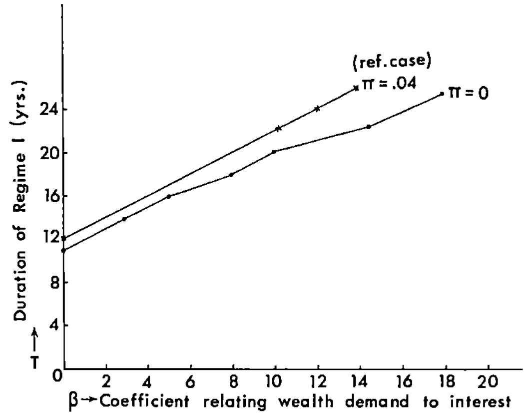


FIGURE 3. EFFECT OF INTEREST SENSITIVITY OF SAVINGS ON DURATION OF REGIME I (POSITIVE GROSS INVESTMENT) FOR TWO INFLATION TARGETS, .04 AND 0.

Note: β's higher than plotted points prolong Regime I more than 25 years.

I reiterate that these simulations are not predictions. They are designed only to illustrate why remedial policies should and will be adopted, and thus to make concrete the vague forebodings about runaway government debt.

Work Incentives in the AFDC System: An Analysis of the 1981 Reforms

By ROBERT MOFFITT*

The road to welfare reform increasingly appears to be one of the rockier paths the United States has traversed. Indeed, by all outward appearances it is not even clear whether the current path runs uphill or downhill. As far as work incentives in the welfare system are concerned, economists of all political persuasions, from Milton Friedman to James Tobin, have agreed that the uphill direction is that which leads to lower tax rates (i.e., lower benefit-reduction rates). However, as with free trade, the near unanimity of opinion among economists has, strangely, only occasionally persuaded a majority of the nation's representatives to vote to go uphill. The legislative history so illustrates. From 1935, when the Aid to Families with Dependent Children (AFDC) program was enacted, to the 1967 Social Security Amendments, the tax rate in the program was 100 percent—that is, benefits were reduced by one dollar for every extra dollar earned. With the 1967 Amendments, Congress lowered the tax rate to 67 percent, and, in the heady atmosphere of the 1960's, it was expected that further progress in this direction would be made and that the tax rate would be lowered further. Indeed, the Family Assistance Plan subsequently proposed by President Nixon would have lowered tax rates; however, the legislation passed the House but not the Senate. The Ford Administration considered welfare reform proposals internally but never proposed legislation, while the Carter Administration proposed a massive welfare reform plan that met with no legislative success. Welfare reform was finally achieved in 1981 when the Omnibus Budget Reconciliation Act (OBRA) was enacted. But OBRA *increased* the tax rate back to 100 percent, the level prevailing prior to 1967. In retrospect, it appears that 1967 marked the end, not the beginning, of legislative progress on work incentives in the welfare system.

In this paper I shall report the results of recent research that complicates the issue considerably by questioning whether lower tax rates do in fact provide work incentives. The findings themselves seesaw not unlike the path of welfare reform itself. First, on a theoretical basis, it appears that lower tax rates in a welfare program do not necessarily increase labor supply in the low-income population as a whole, contrary to the conventional wisdom. In fact, it also appears that members of the Reagan Administration were aware of this all along, well in advance of the economics profession. This theoretical ambiguity has fairly fundamental implications for the work-incentive issue in welfare reform. Second, nevertheless, the empirical resolution of the ambiguity provided by existing econometric estimates in the labor supply literature and by estimates of the effect of AFDC on labor supply indicates that a lower tax rate would indeed increase labor supply in the low-income population as a whole, and that a higher tax rate would decrease it. Thus the conventional wisdom is correct even though based upon an incorrect

†*Discussants*: Henry Aaron, The Brookings Institution and University of Maryland; Harold Watts, Columbia University; Edward Gramlich, University of Michigan.

*Brown University, Providence, RI 02912. Comments from Tom Fraker and Harold Watts are appreciated. I acknowledge financial support for much of the research underlying this paper from the Department of Health and Human Services through a grant to the Institute for Research on Poverty at the University of Wisconsin. All opinions are my own.

theoretical model. Third, given this, one would naturally expect that any evaluation of the 1981 OBRA legislation to detect disincentives of the tax rate increase. On the contrary, virtually all of the studies of OBRA have found no work disincentives whatsoever. There has been considerable comment and puzzlement over these findings and over their uniformity. Fourth, although the evidence is quite tentative, it appears that those studies may have simply stopped too soon. The long-run response to OBRA shows more evidence of work disincentives.

I. The Theoretical Ambiguity and the Evidence

The discussion in this section is drawn from two of my studies on work incentives in welfare systems (1985a, c). The theoretical ambiguity, first shown explicitly by F. Levy (1979), arises because a reduction in the tax rate in a transfer program raises the break-even level in the program—that is, the level below which income must fall for a household to be eligible for benefits. Consequently, while a tax rate reduction may induce some initial recipients to increase their hours of work or to join the labor force, it draws some individuals into the program who had not participated initially. Some of these individuals may be made newly eligible automatically, while others may be induced by the increased generosity of the program and the higher break-even level to reduce their hours of work so as to become eligible. These new recipients will reduce their labor supply when joining the program. The net effect of the tax rate reduction on overall labor supply thus depends upon the sizes of the reductions of the two groups—those initially on the program and those newly drawn into it—and on their relative numbers. More generally, it depends upon the size of labor supply elasticities and the distribution of income.

The opposite occurs for a tax rate increase such as that provided by OBRA. While those who remain recipients will reduce their work effort, those who leave the program will increase it. The net effect depends upon labor supply elasticities and the income distribution.

It should be pointed out that the guarantee level is held constant in these examples. If the guarantee were reduced at the same time as the tax rate so as to hold program expenditures constant, the net effect on labor supply would still be ambiguous but would more likely be positive. If the tax-rate reduction were fully income compensated—that is, if the guarantee were lowered so as to leave every individual in the population on his or her initial indifference curve—the net labor supply effect would again be more likely positive, though still ambiguous because the substitution effects of new recipients are negative. The above model is based instead upon the assumption that there is a donor population that wishes to increase its aggregate transfer to the poor and would like to do so in a way that provides work incentives.

For the same reason, transfer programs that create notches—points where benefits are suddenly terminated by a small increment in income and where tax rates are consequently greater than 100 percent—do not necessarily provide work disincentives.

Since the ambiguity can only be resolved empirically, what does the available evidence have to say? Perhaps surprisingly, despite an enormous amount of research on labor supply over the past 15 years, there is little that is focused directly on this question. It might be thought that direct evidence would be available from the evaluations of the 1967 Social Security Amendments, the most fundamental change in work incentives in the AFDC program up to that time. Unfortunately, the evaluations that were conducted were rather primitive and of little value. For example, most simply compared hours of work of AFDC recipients before and after the tax rate was reduced. It was generally found that hours of work had increased, but this may simply have been a result of the increase in the break-even level. More direct evidence is available from the negative income tax experiments by comparison of mean hours of work in the control group and in the experimental groups with different tax rates. The evidence from the experiments is mixed, for tax rate effects were sometimes positive and sometimes negative. Other explanations for

the experimental findings could be adduced, but from the perspective of the present issue, it is fair to say that the findings do not particularly suggest that hours of work and tax rates are negatively related.

The wealth of estimates of income and substitution elasticities from the nonexperimental labor supply literature (and many from the experimental literature as well) are not of direct use for the issue at hand because the model described above involves a nonlinear budget constraint. The elasticities in the literature are directly useful only for predicting shifts in linear constraints. However, the elasticities implicitly define an indifference map and so can be used to simulate the responsiveness of hours of work to shifts in nonlinear constraints as well. When a range of elasticities is drawn from the literature and used to stimulate the effect of transfer-program tax rates in this way, the results show that hours of work and tax rates are generally negatively correlated for female heads but not always so for men and married women (see my 1985a paper). For married women, for example, the simulations suggest that 100 percent tax rates may minimize work disincentives. However, for female heads, the group relevant to the AFDC program, the results indicate that at current guarantee levels and tax rates a 50 percent change in the tax rate would induce a change in average weekly hours of work of from .58 to 1.12 in the opposite direction. These figures are weighted averages over the entire female-head population, both recipients and nonrecipients, as well as over workers and nonworkers. Since about 50 percent of female heads in the population are in the AFDC program, the corresponding figures for recipients are approximately 1.16 and 2.24 since nonrecipients do not respond to tax rate changes on the margin. Since only 20 percent of AFDC recipients work, and since only workers can respond on the margin, the effects for AFDC working recipients would be five times these figures. Thus the magnitude of the potential response to work incentives does not appear to be small.

Since these estimates are obtained only by simulation, and since they rely on income and substitution elasticities that are based only occasionally upon the analysis of transfer programs of any kind, it would seem desirable to have more direct evidence on the correlation of tax rates and hours of work in the AFDC program. I have recently completed a study (1985c) that estimates partial correlations between the hours of work of female heads in the United States and tax rates in the AFDC program. Estimates from cross sections of states as well as from time-series data over the period 1967–82 were obtained. Somewhat surprisingly, given the fragility of many of the findings in this area of research, the results were very close to those given by the simulations. For example, a 50 percent change in the tax rate was estimated to induce weekly hours changes of from .76 to 1.59, quite similar to those just reported above. The fairly close correspondence of these two independent sources of estimates suggests strongly that tax rates in the AFDC program and labor supply are indeed negatively correlated among female heads.

II. The Effects of the 1981 Reforms

According to Levy (1978), the California Welfare Reform Act of 1971, enacted while Ronald Reagan was governor of California and Robert Carleson his advisor, effectively increased tax rates in the California AFDC program. Levy argues that the governor and his advisors were convinced that the tax rate reduction enacted by the 1967 Social Security Amendments had not provided work incentives, but instead had just increased eligibility and drawn more women onto the rolls —in short, the possibility suggested in the last section. Although the empirical evidence just discussed seems not to substantiate this, it provides an explanation for the 1981 OBRA legislation, which increased nominal tax rates in the program back to 100 percent. The legislation was complex and the changes in the AFDC program were more complicated than has been described here (see my 1985b paper), but an increase in the effective tax rate was the primary work-related change in the legislation.

Immediately subsequent to the legislation a number of studies were initiated to de-

termine whether the legislation would generate work disincentives. (The studies are summarized in my 1985b paper.) Each study followed a group of working AFDC recipients who were on the rolls just prior to the implementation of OBRA (in December 1981–February 1982) for several months afterward. When OBRA was implemented, many women were cut from the rolls and others were given reduced benefits. In each study the number of women returning to the rolls as nonworkers, or remaining on the rolls but at reduced levels of work effort, were compared to "normal" numbers of such movements prior to OBRA. Note that the effect of OBRA on the work effort of women leaving the AFDC rolls and not returning is not obtained, which must be regarded as a design flaw. Nevertheless, the surprising aspect of the findings is that all studies found no detectable work disincentives of the legislation. In every case the number of women returning to the rolls was no greater than it had been prior to OBRA.

There are a number of other issues surrounding these studies which I have no space to discuss, but one possible resolution of the conflict between these findings and those reported in the last section is provided by tabulations I have constructed, shown in Table 1. The table shows trends in hours of work of female heads in the United States from 1968 to 1981 (pre-OBRA) and from 1982 to 1985 (post-OBRA), along with the unemployment rate in each observation year. The data indicate that weekly hours of work fell by 1.1 hours from 1981 to 1982, but that the unemployment rate rose by 2.1 percent simultaneously. On the basis of the sizes of cyclical movements in the table prior to 1981, the 1.1 hours decline was only barely greater than that which would have been expected from the increase in the unemployment rate. But, in 1983, hours of work fell by 1.4 per week even though the unemployment rate declined. In addition, by 1984 hours of work had risen only to 19.2 per week, 1.2 hours below the 1981 level despite a return of the unemployment rate to its 1981 level. From 1984 to 1985, normal cyclical behavior seems to have returned.

TABLE 1—TRENDS IN HOURS OF WORK OF FEMALE HEADS IN THE UNITED STATES 1968–85

	Mean Hours of Work per Week[a]	Unemployment Rate
1968	18.8	3.6
1969	18.6	3.5
1971	17.6	5.9
1973	17.7	4.9
1975	17.2	8.5
1977	18.2	7.0
1979	20.6	5.8
1981	20.4	7.6
1982	19.3	9.7
1983	17.9	9.6
1984	19.2	7.5
1985	19.4	7.2[b]

Source: Tabulations from the *Current Population Survey.*
[a]In March of each year.
[b]Estimated.

The OBRA studies thus appear to have stopped too soon—they generally followed recipients only through 1982 or the first few months of 1983. Negative responses appear not to have begun until 1983 and 1984. Indeed, when simple hours-of-work regressions are estimated on the data in Table 1 controlling for trend and cycle, the 1983–84 values are estimated to be .70 to .90 hours per week below trend (see my 1984d paper). Since OBRA raised the effective tax rate in AFDC by about .46 (T. Fraker et al., 1985), these crude disincentive estimates can be seen to fall into the range of those that would be predicted by the econometric estimates discussed in the previous section.

There are specific reasons for suspecting that the OBRA effect would occur with a lag. The 100 percent tax rate is imposed only after four months of consecutive earnings by an AFDC recipient. Since many women were not working at the time of the implementation of OBRA, and since labor force behavior among AFDC recipients is intermittent, it would take some time for a four-month spell of earnings to occur for many women. More generally, women who were terminated from the rolls may have attempted to stay off AFDC initially, but may have come back onto the rolls after the sickness of a child or some other random event.

Thus, while the evidence in Table 1 is fairly weak by its nature, it does suggest that OBRA may have had the work disincentive effects predicted by the econometric estimates. Only additional research at a micro level can confirm this suggestion.

III. Conclusion

The weight of the evidence discussed here implies that higher transfer-program tax rates such as those resulting from OBRA will reduce work effort among the female-head population, and that lower tax rates would increase it. In conclusion I should like to make two additional points to put this finding in a slightly broader context. First, it should be stressed that tax rates in transfer programs should not be set solely on the basis of their work disincentive effects. Obviously, a transfer program per se has greater work disincentives than no program at all. Instead, the tax rate should be set by balancing the work-disincentive issue against cost considerations and against normative goals for redistribution. For example, a given aggregate transfer is distributed differently by different tax-and-guarantee combinations, and it may be that one wishes to transfer some resources to earners—the "working poor"—per se, independent of work disincentive effects. Second, as mentioned in the paper, the results here do not necessarily extend to men and women in marital unions. If a comprehensive transfer system were extended to those groups, the econometric evidence discussed here suggests that high tax rates may minimize work disincentives. Again, of course, other considerations should play a role in setting the tax rate for those groups as well.

REFERENCES

Fraker, T., Moffitt, R. and Wolf, D., "Effective Tax Rates and Guarantees in the AFDC Program, 1967–1982," *Journal of Human Resources*, Spring 1985, *20*, 264–77.

Levy, F., "What Ronald Reagan Can Teach the United States About Welfare Reform," in D. Burnham and M. Weinberg, eds., *American Politics and Public Policy*, Cambridge: MIT Press, 1978.

______, "The Labor Supply of Female Heads, or AFDC Work Incentives Don't Work Too Well," *Journal of Human Resources*, Winter 1979, *14*, 76–97.

Moffitt, R., (1985a) "A Problem with the Negative Income Tax," *Economics Letters*, 1985, *17*, 261–65.

______, (1985b) "Evaluating the Effects of Changes in AFDC: Methodological Issues and Challenges," *Journal of Policy Analysis and Management*, Summer 1985, *4*, 537–53.

______, (1985c) "Work Incentives in Transfer Programs (Revisited)," mimeo., Brown University, 1985.

______, (1985d) "A Note on the Effect of the 1981 Federal AFDC Legislation on Work Effort," mimeo., Brown University, 1985.

Initial Findings from the Demonstration of State Work/Welfare Initiatives

By DANIEL FRIEDLANDER, BARBARA GOLDMAN, JUDITH GUERON, AND DAVID LONG*

The problem of providing work incentives along with adequate income support to poor families under the federally supported Aid to Families with Dependent Children (AFDC) program has been addressed in two ways. One approach involves changes in the rules for determining welfare eligibility and benefit amount in order to increase the financial incentives for choosing work over welfare. The effect of these voluntary work incentives was tested in the early 1970's in a series of negative income tax experiments utilizing random assignment, which indicated that reductions in the marginal tax rate on earnings for female family heads could prove costly without increasing labor supply much (see, for example, Michael Keeley et al., 1978, and Robert Moffitt, 1979). A second approach would condition AFDC receipt on the fulfillment of an obligation to take a job, search for work, or participate in education or training activities designed to prepare an individual for work. This approach to work is mandatory in that (a) the income entitlement is reduced to sanction individuals who do not comply, and (b) services and suasion, perhaps jobs also, are supplied to induce work directly. Prior to 1982, this approach —initially embodied to a limited degree in the WIN program—was not evaluated with an experimental design but only by less rigorous methods.

Rising popular sentiment that women who head families receiving public assistance have an obligation to work to support their children has led to renewed interest in the mandatory approach. In the last 4 years, building on WIN rules and the added flexibility given them by the 1981 Omnibus Budget Reconciliation Act (OBRA), a majority of states have sought to strengthen work and participation requirements and to broaden the mix of employment and training activities. This paper reports preliminary results from a 5-year, eight-state series of social experiments—the Demonstration of State Work/Welfare Initiatives—that investigate the feasibility, impact, and incremental cost of new state employment strategies for AFDCs.

The research designs call for the random assignment of AFDC applicants or recipients to experimental treatment and control groups—E and C groups. Three of these demonstration programs are located in large urban centers (San Diego, Baltimore, and Chicago); the others are multicounty programs that span urban and rural areas (Arkansas, West Virginia, Virginia, Maine, and New Jersey). This paper presents the first experimental findings for mandatory programs and covers the three demonstration areas with initial results now available: San Diego (Goldman et al., 1985), Baltimore (Friedlander et al., 1985a), and part of Arkansas (the southern half of Little Rock and the city of Pine Bluff; Friedlander et al., 1985b).

I. Mandatory Program Participation

Individual data on program participation were collected from various sources, principally automated state tracking systems. These data were used to address the first question of interest: What kinds of services were provided and how broadly were enrollees covered by a "mandate" to work or to participate in

*Manpower Demonstration Research Corporation, 3 Park Avenue New York, NY 10016. The research summarized in this paper was supported by the Ford Foundation, the Winthrop Rockefeller Foundation, and the participating states. However, the findings and interpretations do not necessarily represent the official position or policy of the funders.

assigned activities? Table 1 offers a statistical description of the experience of a cohort of controls and experimentals entering the three programs as well as information pertinent to the interpretation of that experience. For example, statutory grant maximums, which are based on individual state standards of need, vary widely among states. Low benefit levels not only reduce the attractiveness of AFDC receipt relative to low-wage employment, but also increase the likelihood of case closure when employment is obtained

San Diego and Baltimore required participation from women with no preschool children—the usual "WIN-mandatory" group, which comprises roughly one-third of AFDC case heads. Arkansas extended mandatory status to mothers whose youngest child was age 3 or older. Random assignment was performed at the point when individuals were told they had to enroll in the program, but the demonstrations differ in groups targeted for enrollment, scale, kinds of activity and degree of mandatoriness, and services to controls. In San Diego, applicants for AFDC were randomly assigned at the point of application. Experimentals were required to enroll in a job-search component, which was followed by 13 weeks of unpaid work experience—"workfare"—for those who did not find jobs. (The short-run experience of a second experimental group assigned to job search only was similar and to conserve space is not discussed in this paper.) Controls in San Diego were assigned to WIN, where they received a very limited amount of employment and training services.

Arkansas randomly assigned both applicants and those current recipients whose status changed to mandatory during the demonstration. For the experimental group, the program provided job search, followed for a few individuals by short-term unpaid work experience. Arkansas controls, unlike those in San Diego, were a no-services group, since the experimental program replaced WIN.

In Baltimore, the program worked with applicants and newly mandatory recipients but was oriented toward longer-term employability development as well as immediate placement. Unlike the other two programs, its activities were not in fixed sequence: clients, with the guidance and consent of their caseworkers, were required to choose from among job search, work experience, education, and training. The program restricted enrollment to 1000 per year and maintained a higher level of funding, per experimental, than the other two programs. Controls received regular WIN services, which consisted of some short-term help finding jobs and occasional training.

Table 1 reports two measures of administrative outcomes: participation and cover-

TABLE 1—WELFARE TURNOVER AND RATES OF PROGRAM PARTICIPATION AND COVERAGE

	San Diego	Arkansas	Baltimore
Maximum Grant			
Family of 3[a]	\$526	\$140	\$295
Program:			
Clients/Yr.	unlim	unlim	1000
Sequence	fixed	fixed	discretionary
Services to Controls	WIN	none	WIN
Impact Sample:			
E group	883	554	1331
C Group	538	565	1372
Percent of Controls			
AFDC Applicant	100.0	59.3	51.2
Received AFDC:[b]			
Ever During Follow-Up	85.0	78.6	95.1
Quarter 4	48.6	63.7	73.2
Percent of Experimental Registrants[c,d]			
Participated[e]	54.7	38.0	52.8
Not Covered	11.9	23.7	23.8
Sanctioned	9.9	4.3	rare
Percent of Experimental Participants[c]			
Component Mix:			
Independent Job Search	–	61.3	54.7
Job Club	94.5	71.8	
Work Exper.	28.2	7.6	39.8
Educ-Train.	–	–	38.6

[a]Parent, two children. October 1, 1982.
[b]Arkansas receipt is based on a sample smaller than its impact sample.
[c]9-month follow up for San Diego and Arkansas; 12-month for Baltimore.
[d]Percent of experimentals who registered were 86.1, 100.0, and 84.9 for San Diego, Arkansas, and Baltimore, respectively.
[e]Attended at least one activity.

age. As the table shows, 54.7 percent of those who registered with the program in San Diego attended at least one activity within 9 months after random assignment. This *participation rate* meets the intent of program planners and exceeds that of many comparable past programs. Yet, the measure has limitations as an indicator of program performance. A normal attrition from the group of registrants available to start activities occurs when AFDC applications are denied, when families leave the welfare rolls, and when individuals receiving AFDC are determined to have personal or family circumstances that exempt them from a participation requirement. Substantial movement out of the welfare system, even without special program services, is evident in Table 1 for control-group applicants in San Diego: only 85.0 percent were approved for AFDC and only 48.6 percent were still on the rolls at the fourth follow-up quarter. This pronounced natural caseload turnover means that many individuals in any registrant cohort would be out of the program before their participation was scheduled to begin, making it impossible for even a mandatory program to approach 100 percent participation among enrollees.

Consequently, although participation rates are useful in describing level and mix of activity, the *coverage rate* better captures the "reach" of a program's mandate. This rate counts not only participants but also nonparticipants who were sanctioned by the program for noncompliance, or left welfare, or for other reasons were deregistered from the program and therefore left its purview. The coverage rate thus gives the percent of registrants—9 months after random assignment for San Diego—that either had participated or were penalized for noncompliance or were out of the program altogether. Table 1 reports one hundred minus this rate, which is the proportion still enrolled but not yet worked with. For San Diego, coverage was reasonably complete: the local agency did implement a short-term participation requirement. The 11.9 percent of registrants not covered at 9 months consisted mostly of persons deemed too difficult to work with, owing largely to language barriers and health exemptions granted by local office staff. It also included individuals who were employed but still on welfare and required to remain registered with the program. Cross-state comparisons of program coverage are, unfortunately, problematic, owing in part to differences in criteria for exemption of enrollees from participation. Also, in programs where recipients were enrolled along with applicants, the latter group had much higher coverage rates than the former.

In Arkansas, some 38.0 percent of registrants participated in the program within 9 months after random assignment, and 23.7 percent were left uncovered. In part, lower coverage than in San Diego reflects much lower caseload turnover among AFDC recipients, who were not present in the San Diego program. Administrative constraints and other factors also limited the extent to which Arkansas could operate a mandatory participation program.

The Baltimore program achieved a participation rate of 52.8 percent; in contrast, fewer than 4 percent of controls participated in a major WIN component. As in Arkansas, the 23.8 percent not covered in Baltimore is partly associated with the intermixing of recipients and applicants. It also reflects Baltimore's lesser emphasis on mandatory participation compared to San Diego as well as the fact that, consistent with the program's goal of longer-term employability development, individuals enrolled in nonprogram schooling could be deferred.

Table 1 also lists activities in the programs' component mixes in order of ascending cost and intensity. Sanctioning rates and component mix statistics are in line with the different state philosophies and resource commitments. Absenteeism was frequent, but in all states persuasion was more widely used than sanctioning in securing compliance. A major survey of San Diego experimentals, as well as small samples of work experience participants in all three programs, revealed that most felt program participation requirements were fair, perhaps because local staff tended to portray participation as an opportunity.

II. Program Costs

State and county expenditure data were used to estimate program cost per experi-

mental and per control, including nonparticipants and program participants alike. Cost estimates in Table 2 give the reader a sense of the relative overall intensity of the program treatment across states and between experimentals and controls within a state. The per experimental cost of San Diego's program was less than one-half the $1050 cost of Baltimore's program; Arkansas' program cost about one-sixth as much as Baltimore's. Costs of the job search and work experience components per enrollment week were similar in San Diego and Baltimore but were lower in Arkansas due largely to less intensive staff involvement. San Diego spent more on achieving compliance with its participation requirement (which entailed monitoring, sanctioning, and registrant follow up), while Baltimore devoted considerably more resources to education and training and on direct payments to clients (largely stipends and child care payments).

These figures represent the gross costs of the programs, not their net costs. In estimating the net cost, one must take account of the costs of serving the control group—which were negligible in Arkansas where no services were provided, about $80 per control in San Diego, and about $141 per control in Baltimore. The complete benefit-cost analyses, of course, also consider projected savings in AFDC payments generated by the programs and other effects beyond the scope of this paper.

TABLE 2—SHORT-TERM EXPERIMENTAL-CONTROL DIFFERENCES

	E-Group Mean	*C*-Group Mean	*E-C*
Program Cost			
San Diego	$485	$80	+$405
Arkansas	$165	$7	+$158
Baltimore	$1050	$141	+$909
Quarter 3: Percent Ever Employed			
San Diego	37.6	28.7	+9.0
Arkansas	15.2	12.2	+3.1
Baltimore	32.4	27.9	+4.5
Quarter 3: Percent Received AFDC			
San Diego	54.2	58.7	−4.5
Arkansas	56.8	63.8	−6.9
Baltimore	77.4	78.2	−0.8
Quarter 3: AFDC Dollar Amount			
San Diego	$608	$668	−$60
Arkansas	$246	$289	−$43
Baltimore	$594	$593	+$0

Notes: Program cost is per *E* or *C*. Employment and welfare receipt are for quarter 3 of follow up. Discrepancies in sums and differences are due to rounding.

All *E-C* differences in employment and welfare receipt were statistically significant at .10 or above except for welfare impacts in Baltimore.

III. Short-Term Impacts on Employment and AFDC Receipt

The last question to be examined concerns the effect of the programs on the employment and welfare receipt of sample members. WIN administrators have historically judged program performance by the rate at which enrollees find employment (the *placement rate*), and by the number of case closings for enrollees. However, substantial employment and rapid caseload turnover among controls in the absence of special services clearly implies that the usual measures will radically overestimate true program effectiveness.

For this study, program impacts were estimated as experimental-control differences in quarterly employment and welfare receipt, counting as "quarter 1" the quarter in which an individual was randomly assigned. Differentials were regression adjusted to increase statistical precision and control for any dissimilarities in prior employment, welfare history, and demographic characteristics. Quarterly Unemployment Insurance earnings and AFDC payments obtained from computerized state or county administrative records provided the outcome data. Demographic data were obtained during the random assignment interview. Comparison of *all* experimentals—program participants and nonparticipants alike—with *all* controls sidesteps serious problems of selection bias that would arise if the analysis were performed on participants only, and allows for the possibility that some nonparticipants may have become employed or left welfare in order to avoid participation requirements.

Table 2 presents impact estimates for the third quarter of follow-up, the last currently available in all three states. In all cases, employment impacts were evident in the immediate postrandom assignment quarter (quarter 2) and continued throughout the

observation period. The short-term employment impacts of these mandatory programs are comparable to those found in the earlier experimental studies of job search for volunteers.

For San Diego, the first wave of enrollees showed quarterly employment rate increases of 7 to 9 percentage points, with proportionate increases in quarterly earnings. Welfare expenditure reductions were smaller than employment and earnings increases: in the third quarter, experimental-control differences amounted to $-\$60$, a saving of 9 percent relative to benefits paid to controls (with persons not receiving AFDC in a quarter counted as zeroes).

Starting from very low control group employment levels, the program in Arkansas achieved improvements of 3 to 5 percentage points in each quarter. By the third quarter, incidence of welfare receipt had dropped 7 percentage points from 64 to 57 percent; benefit payments were down 15 percent ($43) for the quarter. Also in Arkansas, the extension of mandatory status to mothers whose youngest child was between 3 and 5 years of age yielded employment impacts for this group that were about the same as impacts for the regular WIN-mandatories. Since mothers with the younger children made up 54 percent of the enrollment sample, the findings provide tentative support for the view that broadening mandatory coverage would increase the total program effect on the caseload.

In Baltimore, quarterly employment impacts of 3 to 5 percentage points were observed. Impacts on welfare were small and not statistically significant during this period. There is sketchy evidence that impacts on earnings in Baltimore began to exceed employment gains after five quarters. Such a finding would be consistent with the Baltimore program's stress on long-term employability development and its use of lengthy education and training activities. Longer follow-up is required to investigate this possibility.

Disaggregation of impacts for AFDC subgroups suggests that program effectiveness and optimal component mix may hinge in important ways on the choice of target population. Evidence is accumulating that "less-employable" subgroups experience larger employment impacts. In all three states, experimentals who had no earnings in the year prior to random assignment achieved employment gains over similar controls that at the end of follow-up were more than twice as large as the gains for individuals who did have a recent work record. By the same token, no net employment gains were demonstrated for a large sample of AFDC-U's in San Diego or a small sample in Baltimore. These AFDC-U's, mostly male heads of two-parent households, had stronger records of prior employment than AFDC's, and the AFDC-U control groups exhibited much more of a tendency to increase their employment and reduce their welfare receipt without special program services.

IV. Future Work

The available evidence for these three states indicates that mandatory programs of job search, short-duration unpaid work experience, and training can be implemented for AFDC clients, that they can be operated at relatively low cost, and that they can have some short-run effect on employment. Much work does, however, remain to be done: impacts for the second round of states and further follow up for the first three; computation of benefit-cost estimates; and fuller comparisons across program models and client subgroups. Answers to important questions hinge on the availability of longer-term follow-up data. For example, given the similarity in short-run employment impacts, the shape of impact decay is critical in distinguishing the overall impact and cost-effectiveness of the several program models.

Additional data may provide a better answer to the question, How does a work or participation obligation affect employment among AFDC recipients? While this question is important for *all program eligibles*, it is especially important for those who would actually have been *on AFDC for an extended period of time*. In this regard, the preliminary finding of larger employment gains for individuals without recent work history suggests that other predictors of extended dependence should be tested for association with larger program impact. If such associations are

found to be general in longer-term follow-up, then the full-sample employment impacts reported in this paper may be magnified when adjusted to a base of long-term AFDC recipients.

The finding of substantial natural welfare turnover among controls, particularly for applicants, is consistent with the description of caseload dynamics provided recently by Mary Jo Bane and David Ellwood (1983). High turnover implies that any strategy of immediate "saturation" of entering cohorts with expensive components would entail program outlays for many participants who would have left the welfare system shortly anyway. On the other hand, program operators often argue that delaying services decreases their potential effect because some individuals develop a dependence on welfare that is difficult to overcome. To resolve this issue, further research is required into the relative impacts and costs of service provision that is broad vs. targeted, immediate vs. delayed. In this connection, relationships between subgroup characteristics and net impacts may suggest improved program targeting strategies and clearly deserve further study.

REFERENCES

Bane, Mary Jo and Ellwood, David, "The Dynamics of Dependence: The Routes to Self-Sufficiency," Urban System Research and Engineering, Inc., Cambridge, 1983.

Friedlander, Daniel et al., (1985a) *Maryland: Final Report on the Employment Initiatives Evaluation*, New York: Manpower Demonstration Research Corporation, 1985.

______, (1985b) *Arkansas: Final Report on the WORK Program in Two Counties*, New York: Manpower Demonstration Research Corporation, 1985.

Goldman, Barbara et al., *Findings from the San Diego Job Search and Work Experience Demonstration*, New York: Manpower Demonstration Research Corporation, 1985.

Keeley, Michael C. et al., "The Labor Supply Effects and Costs of Alternative Negative Income Tax Programs," *Journal of Human Resources*, Winter 1978, *13*, 3–36.

Moffitt, Robert A., "The Labor Supply Response in the Gary Experiment," *Journal of Human Resources*, Fall 1979, *14*, 477–87.

An Evaluation of the Effect of Cashing Out Food Stamps on Food Expenditures

By THOMAS FRAKER, BARBARA DEVANEY, AND EDWARD CAVIN*

The Omnibus Budget Reconciliation Act of 1981 mandated that the Commonwealth of Puerto Rico's participation in the U.S. Food Stamp Program (FSP) be replaced by an annual $825 million block grant to provide food assistance for needy persons. The Commonwealth responded by designing the Nutrition Assistance Program (NAP), which was implemented on July 1, 1982. The NAP differs from the former FSP in several respects. First, and of primary importance, food coupons have been replaced by monthly checks that are freely negotiable for currency. In addition, the switch to NAP included reductions in eligibility limits and benefit standards in order to bring program costs into line with the legislatively reduced funding level of the block grant.

The legislative history of NAP reflects congressional concern that cash assistance would weaken the already tenuous link between program benefits and food consumption. Because of this concern and because some members of Congress felt that the cash program in Puerto Rico was a first step to cashing out the FSP in the United States, one provision of the Omnibus Budget Reconciliation Act of 1982 required Puerto Rico to return to a noncash system of food assistance. Implementation of this requirement was delayed until July 31, 1985 by subsequent legislation, which also mandated that the Secretary of Agriculture conduct an evaluation of NAP. This paper presents findings from that congressionally mandated evaluation. Specifically, the central policy issue addressed by this paper is whether the replacement of food coupons with cash assistance in Puerto Rico has resulted in reductions in food expenditures by participating households. Harold Beebout et al. (1985a, b) present comprehensive results from the evaluation of the effects of NAP not only on food expenditures but also on the nutritional adequacy of diets, administrative costs, and program fraud and error.

The two data sets that were the basis for the NAP evaluation are described in the following section. The statistical model that was used to estimate the effects of cash coupon benefits on food expenditures is presented in Section II, followed in Section III by a discussion of the results of the empirical analysis. Conclusions drawn from the empirical findings are presented in the final section, along with a discussion of NAP's legislative outcome.

I. The Data

The data used in this analysis are from two household food use surveys conducted in Puerto Rico. The first is the 1977 Puerto Rico Supplement to the *Nationwide Food Consumption Survey*, which was fielded between July and December 1977 when the former FSP was operating in Puerto Rico. The second survey was fielded in Puerto Rico between July and December 1984, after NAP had been in effect for 2 years, and is called the 1984 Puerto Rico *Household Food Consumption Survey*. Both survey samples were representative of Puerto Rico's population and were identical in terms of the data collection methodology. The 1977 analysis sample includes 2,940 households, while the 1984 analysis sample is somewhat smaller,

*Fraker: Mathematica Policy Research, Inc., 600 Maryland Avenue, S.W., Suite 550, Washington, D.C. 20024. Devaney and Cavin: Mathematica Policy Research, Inc., P.O. Box 2393, Princeton, NJ 08540. This research was funded by the Food and Nutrition Service of the U.S. Department of Agriculture under Contract No. FNS 53-3198-4-63. The opinions expressed herein as well as any errors are our own and not of the sponsoring agency.

totaling 2,423 households. (See Beebout et al., 1985b, for full data sources.)

These two surveys provide detailed information on household food use. Household food use refers to food and beverages used from household food supplies during the seven days preceding the interview. Food used that was purchased with cash, credit, or food stamps and food used that was home produced or received as a gift or payment for work all are considered components of total household food use.

It is important to note that household food use is not equivalent to food intake by individuals in the household. Food intake refers to food actually eaten and may be substantially less than food used. The difference between the amount of food that is used from the household food supplies and actual food intake can be attributed to food waste or loss.

A. *Measures of Food Expenditures*

Two measures of food expenditures are analyzed in this paper. The first is total food expenditures, defined as the sum of the money value of food used at home, the amount spent on meals and snacks away from home, and the subsidy value of school lunches and school breakfasts. The second is the money value of food used at home, which is the money value of food used from household food supplies by household members and guests. It is derived from the reported quantities of the individual food items used by the household during the seven-day period preceding the interview. The money value of food used is obtained by multiplying the quantity (in pounds) of each food item used by its respondent-reported price per pound. Food not purchased directly by the household (i.e., home-produced food or food received as a gift or pay) is valued at the average price per pound for that food item that was paid by the survey households reporting its purchase and use. The total money value of food used at home is obtained by summing the money values of the individual food items. All expenditure measures discussed in this paper are expressed in constant (1984) dollars.

B. *Measures of Household Composition*

A consistent finding of previous research based on food use data similar to the data analyzed for this evaluation is that household size and composition have important effects on food expenditures. Larger households and households with certain types of members (for example, teen-aged males) are found to have higher food expenditures than households of other sizes and/or composition.

Two basic measures of household composition are used in this paper: household size in adult-male-equivalent persons; and household size in equivalent nutrition units. Household size in adult-male-equivalent persons is actual household size adjusted for the age and sex of the household members. The adjustment procedure weights each household member by the nutritional requirements of that member relative to the nutritional requirements of an adult male age 23–50, where the nutritional requirements for this analysis are based on the 1980 Recommended Dietary Allowances for food energy. The sum of these weights gives household size in adult-male-equivalent persons.

Household size in equivalent nutrition units is the number of adult equivalent males in the household eating meals from the household food supplies. It adjusts actual household size for both the age-sex composition of the family members and the proportion of meals eaten at home. This measure of household size is particularly important for the analysis of the money value of food used at home, as discussed below.

II. The Model

The basic model of food expenditures and program participation estimated for this evaluation is the following:

$$F_i = X_i\beta + \alpha B_i + \varepsilon_i \tag{1}$$

$$P_i = 1 \quad \text{if} \quad Z_i\delta + u_i \geq 0 \tag{2}$$

$$= 0 \quad \text{if} \quad Z_i\delta + u_i < 0,$$

where F_i is a measure of food expenditures

of the ith household, X_i is a vector of household characteristics affecting food expenditures, P_i is a dichotomous variable denoting program participation, B_i is the food assistance benefit amount (zero for nonparticipants), Z_i is a vector of household characteristics (which may or may not contain elements of X_i) that influence the FSP or NAP participation decision of program eligibles, and ε_i and u_i are random disturbance terms. The random disturbance terms are assumed to have a bivariate normal distribution, with a nonzero correlation.

The primary objective of the analysis of food expenditures is to obtain unbiased estimates of α, the marginal effect of food assistance benefits on food expenditures, based on the before- and after-NAP data. The effect of cash issuance of food assistance benefits is gauged by the change in coefficients; if cash issuance reduced food expenditures, the 1984 value of α should be less than the 1977 value.[1]

Most previous analyses of the effect of food stamps on food expenditures (with the notable exception of a study by Jaing-Shing Chen, 1983) have obtained estimates of α simply by estimating equations similar to equation (1) without reference to the program participation decision. A potential problem with such estimates of α is that FSP or NAP participants may be self-selected groups of households that have greater food expenditures than otherwise similar eligible nonparticipants even in the absence of a food assistance program. Failure to recognize the interdependence of the food expenditure and program participation equations therefore is likely to result in biased estimates of α.

A full-information maximum likelihood (*FIML*) procedure was used to obtain separate estimates of the model for 1977 and 1984. By taking into account the correlation between the disturbance terms in the expenditure and participation equations, *FIML* produces estimates of the expenditure equation that are free of sample selection bias. The *FIML* estimates of the participation equation can be interpreted as if they were probit estimates.

III. Empirical Results

For the purpose of evaluating the effect of the cash issuance of food assistance benefits on food expenditures, the model presented above was estimated on data from both the 1977 Puerto Rico Supplement to the *Nationwide Food Consumption Survey* and the 1984 Puerto Rico *Household Food Consumption Survey*. Two different specifications of the dependent food expenditure variable were used: total household food expenditures per adult male equivalent and the money value of food used at home per equivalent nutrition unit. The household size and composition scale differs for the two food expenditure variables because, for the analysis of the money value of food used at home, it is important to standardize by the number of meals eaten at home. Otherwise, if the number of meals away from home were greater in 1984 than in 1977, then NAP would appear to have reduced the money value of food used at home per adult male equivalent regardless of whether any change occurred in the money value of food used at home per meal at home.

No previous study has examined directly the effect of cash food assistance on household food expenditures; however, many estimates of the effects of food stamps and money income on food expenditures are available from analyses of U.S. data. Most previous estimates of the marginal propensity to consume food (*MPC*) out of money income are between .05 and .10, implying

[1]Actually, the 1977 data were collected before another major policy change to the FSP in Puerto Rico—the elimination of the purchase requirement (*EPR*) in 1979. As both *EPR* and the implementation of NAP occurred between 1977 and 1984, a comparison of the 1977 and 1984 coefficients on food assistance benefits to determine the effect of cash issuance may be misleading since any differences in the coefficients may be due in part to *EPR*. As part of this evaluation, an econometric analysis of the 1977 data was conducted to investigate the effects of the purchase requirement on food expenditures. The results of this analysis indicated no impact of the purchase requirement and, hence, no impact of *EPR* on food expenditures. Thus, the comparison of the 1977 and 1984 coefficients on food assistance benefits provides an unbiased assessment of the impact of cash issuance on food expenditures (Beebout et al., 1985b).

that household food expenditures increase by 5 to 10 cents in response to an additional dollar of money income. In contrast, most estimates of the *MPC* out of food stamps are between .20 and .45. However, there are several reasons why these estimates for the United States are of little value in assessing the effect of replacement of food stamps with cash food assistance in Puerto Rico: 1) the great disparity in per capita income between Puerto Rico and the United States, as well as cultural differences that are likely to influence food consumption, suggests the existence of a substantial difference in the *MPC* out of money income; 2) an estimate of the *MPC* out of money income may be a poor proxy for an estimate of the *MPC* out of cash food assistance; and 3) the existence of an active black market for food stamps in pre-NAP Puerto Rico may have contributed to a U.S.-Puerto Rico difference in the *MPC* out of food stamps. While Laura Blanciforti (1983) estimated food expenditure models on Puerto Rico data for the pre-NAP period, her estimates also are of little value in assessing NAP's impact for reason 2 above.

The FIML estimate of the marginal propensity to consume total food out of food stamps (based on the 1977 data) is .21, with a standard error of .05. The corresponding estimate for NAP benefits (based on the 1984 data) is .23, also with a standard error of .05. Both of these estimates are different from zero at the .01 level of significance, but they are not significantly different from each other. The estimates imply that, on average, when Puerto Rico households receive an additional dollar of food assistance benefits, whether in the form of food stamps or cash, they increase their total expenditures on food by slightly more than 20 cents.

When the dependent food expenditure variable is defined to be the money value of food used at home, the estimate of the *MPC* out of food stamps is .27 and the estimate of the *MPC* out of cash food assistance is .21, with equal standard errors of .05. These estimates imply that an additional dollar's worth of food stamps results in an increase in the money value of food used at home that is 6 cents greater than the increase resulting from additional dollar of cash food assistance. This estimated difference is not statistically significant but it is consistent with the fact that food stamps can legally be used only to purchase food for use at home, while cash assistance can be used to purchase not only food to be used at home but also food to be used away from home and nonfood items.

As noted above, eligible households who choose to participate in a food assistance program may have different food expenditures than eligible nonparticipants with similar observed characteristics, irrespective of the amount of assistance. To avoid the bias in estimates of the food expenditure equations that could arise from such differences, these food expenditure equations were estimated simultaneously with FSP and NAP participation equations. The error terms in the expenditure and participation equations were permitted to be correlated. As anticipated, the probabilities of participating in the FSP and NAP were found to increase with the amount of the potential food assistance benefit and to decrease with household income, after controlling for other household characteristics. However, contrary to expectations, no statistically significant correlation was found between the error terms in the food expenditure and program participation equations. This finding indicates that ordinary least squares regression estimates of the food expenditure equations would not be seriously biased by the self-selection of eligible households into food assistance programs.

IV. Conclusions

The results of this evaluation show essentially no difference in the marginal propensities to consume food out of coupons and cash benefits with respect to total food expenditures. For the money value of food used at home, the difference between the *MPC* out of cash benefits and the *MPC* out of coupons is .06, although this estimated difference is not statistically different from zero. These findings of no change in total food expenditures and a small decline in the money value of food used at home suggest that households increased their expenditures

on food away from home as a result of the switch to cash issuance under NAP. However, this substitution of food away from home for food at home is expected to be small, primarily because the estimated reduction in the money value of food used at home per equivalent nutrition unit is both small in magnitude (approximately $.68 per week) and not statistically different from zero.

Congressional staff members were fully briefed on the findings of this evaluation. The response was surprisingly muted, given the strength of some of the original objections to the cash issuance of food assistance benefits. No significant effort was made to dispute the findings and the cash program was permitted to continue beyond its scheduled termination date of July 31, 1985. The 1985 farm bill passed by Congress includes language that rescinds the requirement that Puerto Rico return to a noncash program. However, although Congress now appears reluctant to terminate Puerto Rico's system of cash food assistance, it has not shown much interest in implementing such a system in the United States.

REFERENCES

Beebout, Harold et al., (1985a), *Evaluation of the Nutrition Assistance Program in Puerto Rico*, Vol. I: *Environment, Participation, Administrative Costs, and Program Integrity*, A Report to the U.S. Congress, Washington, D.C., Mathematica Policy Research, Inc., 1985.

______, (1985b) *Evaluation of the Nutrition Assistance Program in Puerto Rico*, Vol. II: *Effects on Food Expenditures and Diet Quality*, A Report to the U.S. Congress, Washington, D.C., Mathematica Policy Research, Inc., 1985.

Blanciforti, Laura, "Food Stamp Program Effects in Puerto Rico," Economics Research Service Staff Report, U.S. Department of Agriculture, 1983.

Chen, Jain-Shing, "Simultaneous Equations Models With Qualitive Dependent Variables: A Food Stamp Program Participation and Food Cost Analysis," unpublished doctoral dissertation, University of Missouri, 1983.

Wage Setting, Unemployment, and Insider-Outsider Relations

By ASSAR LINDBECK AND DENNIS J. SNOWER*

A good conceptual test of any choice-theoretic analysis of persistent involuntary unemployment in free-market economies is to ask whether it can explain why unemployment cannot be eliminated through underbidding. In particular: (a) Why are involuntarily unemployed workers unwilling or unable to gain jobs by underbidding their employed comrades? (b) Why are laid-off workers unwilling or unable to retain their jobs by underbidding? The lower wage bids may be initiated by the firms, the workers, or both in conjunction; "underbidding" is said to take place when the relevant parties accept such bids.

One obvious explanation could be "social norms," according to which underbidding is not an acceptable form of social behavior: "Thou shalt not steal the job of thy comrades by underbidding them," and "Thou shalt not permit job theft from underbidding" are widely accepted social precepts. These imply that existing wages become "fair wages," independent of current demand and supply pressures in the labor market—much as the so-called "wage norms" operate in studies of George Perry (1980), Arthur Okun (1981), and Daniel Mitchell (1985).

However, for an economist, it is natural to try to explore the rationale for such norms when attempting to answer the two questions above. Nowadays the preponderant answer to these questions is contained in the efficiency-wage theories. This paper suggests a logically independent theory: the "insider-outsider" approach. In the efficiency-wage theories, all labor market power rests with the firms, who make the wage and employment decisions under asymmetric information. It is not in the firm's interest to accept the underbidding of involuntarily unemployed workers, because firms use wages as a screening device for productivity. In this case, the unemployment may be understood in terms of a conflict of interest between the firms and the unemployed workers.

By contrast, the insider-outsider approach places some labor market power into the hands of the employees. The crucial assumption is that it is costly to exchange a firm's current, full-fledged employees (the insiders) for unemployed workers (the outsiders), and that the rent associated with this turnover cost can be tapped by the insiders in the process of wage negotiation. Thus wages may be set so that involuntary unemployment results, but the outsiders are nevertheless unable to improve their position through underbidding, because the insiders make underbidding expensive for the firms to accept and disagreeable for the outsiders to pursue. Accordingly, involuntary unemployment arises out of a conflict of interest between the insiders and the outsiders. (See our 1985b paper.) It should be observed that, unlike the efficiency-wage theories, the insider-outsider approach does not assume a direct effect of wages on productivity.

In what follows, we explore the insider-outsider approach by examining how persistent involuntary unemployment can arise under three separate types of cost from insider-outsider turnover: (*i*) the costs of hiring and firing (see our 1986 paper and Robert Solow, 1985); (*ii*) the costs that arise when insiders are prepared to withdraw cooperation from entrants (and thereby reduce the entrants' productivity) or to damage entrants' personal relations with them (and thereby raise the entrants' disutility of work)

*Institute for International Economic Studies, University of Stockholm, S-106 91 Stockholm, Sweden, and Birkbeck College, Department of Economics, 7-15 Gresse Street, London W1P 1PA, England, respectively.

(see our 1985a paper), and (*iii*) the costs implicit in the adverse effect of labor turnover on work effort (see our 1984a paper). We then take a brief look at the implications of the insider-outsider approach for the theory of labor unions.

I. Hiring and Firing Costs

These are perhaps the most conspicuous turnover costs (for example, the expense of implementing mandatory hiring and firing procedures, engaging in litigation, making severance payments) and they take time to incur. Accordingly, it is a convenient simplification to distinguish among three groups of workers: *insiders* (on whom all the hiring costs have been expended and whose dismissal would trigger the full range of firing costs), *entrants* (who are associated only with hiring costs), and *outsider* (who are unemployed and thus require none of the costs above).

For simplicity, imagine each of these groups to be homogeneous. In particular, suppose that all entrants are associated with the same hiring costs and go through a fixed "initiation period," after which they become associated with the same firing costs.

In accordance with the observation that long-term wage contracts (extending over the entire time which employees spend at their jobs) are usually unenforceable, we make the simplifying assumption that wage contracts (for insiders and entrants) last only for a fixed, limited time span, which we set equal to the initiation period. Thus, once an entrant has gone through this period, he has the same job characteristics and bargaining opportunities as an insider; in fact, he becomes an insider.

For the moment, suppose that the insiders are not unionized, so that we can assume the insider wage to be the outcome of an "individualistic" bargaining process between each insider and his firm, whereby the insider takes the wages and employment opportunities of all other workers as given. We require that this bargaining process satisfy two general properties: (*i*) each insider captures some (or all) of the rent inherent in the hiring and firing costs, and (*ii*) the greater this rent, the greater the insider wage.

Then it can be shown that the insider wage will exceed the entrant wage by some positive amount which is not greater than the marginal firing costs. In the same vein, the entrant wage will exceed the outsiders' reservation wage by no more than the marginal hiring costs.

Now consider an economy in which all wages are determined in this way and, at these wages, aggregate labor demand falls short of aggregate labor supply. Is the resulting unemployment "involuntary"?

According to one common definition, involuntary unemployment exists when, at the prevailing wages, workers unsuccessfully seek jobs for which they have the same ability as the current job holders. Yet for our purposes, this definition is too narrow, since insiders, entrants, and outsiders may have different "abilities" to fulfill the available jobs. We can make this idea more precise by dividing the costs of hiring and firing into two categories: (*i*) "indispensable labor costs," without which the act of production could not be performed (for example, screening and search costs in the labor market), and (*ii*) "dispensable labor costs," which are transfers whose abolition would have no intrinsic effect on production (for example, severance payments). Clearly, the indispensable, but not the dispensable, costs should be taken into account in evaluating workers' relative abilities. To deal with unemployment in the presence of ability differences, we provide the following, broader definition of involuntary unemployment: it exists when workers unsuccessfully seek jobs at wages which fall sufficiently below the prevailing wages to compensate the firm for ability differences.

In the context of our economy above, let us measure the ability differences between insiders and outsiders by the differential between their marginal products net of indispensable labor costs. Whenever this ability differential is *less* than the differential between the insider wage and the reservation wage, then the outsiders may be identified as involuntarily unemployed, in the sense that they are arbitrarily exposed to a more restricted opportunity set than the insiders. Moreover, this unemployment will persist whenever the ability differential net of indis-

pensable *and* dispensable labor costs, is *greater* than the wage differential. For, in that event, the firms have no incentive to replace insiders by outsiders.

II. Cooperation and Harassment

Another potentially important reason why insiders may have a stronger bargaining position than outsiders is that the insiders often have considerable latitude in choosing whether to be cooperative with entrants in the process of production or whether to have, or not to have, good personal relations with them. Thus, insiders are able to affect *both* entrants' productivity via work cooperation *and* their disutility of work via unfriendly attitudes, which we simply call "harassment." Firms generally find it impossible to monitor such "cooperation" and harassment activity perfectly and the wage contracts cannot be made contingent on them. (Output-related wage contracts may not obviate this difficulty, because in many cases firms and/or insiders could find them incentive incompatible, too risky, or too costly to monitor, as shown in our 1985a paper.) Under these circumstances, insiders can protect themselves from underbidding by being prepared to withdraw cooperation from the underbidders or to damage their personal relations with them. In other words, the possibilities of pursuing cooperation and harassment generate economic rent which insiders can exploit in wage determination.

To begin with, let us examine the effects of cooperative activities alone. Suppose that wages are determined by the same individualistic bargaining process as above and that insiders can engage in cooperative activities while entrants cannot. Under these circumstances, entrants receive the reservation wage (and thus are not better off than the outsiders). Moreover, it can be shown that the insider wage, generated by the bargaining process, will exceed the reservation wage by some positive amount which is not greater than the insider-entrant marginal product differential, generated by the disparity between insider-insider cooperation and insider-entrant cooperation.

Assuming for simplicity that cooperative activity has no direct utility cost to the insiders, it is in the insiders' interests to make this disparity as large as possible. They do this by cooperating with one another but refusing to cooperate with entrants.

In an economy which runs along these lines, persistent involuntary unemployment may exist in the following sense. The inherent ability difference between an insider (on the one hand) and an entrant or outsider (on the other) stems exclusively from their different individual abilities to provide cooperation to their colleagues. The corresponding ability-related marginal product differential may be evaluated as the amount by which their marginal products would differ under identical external condition of employment, viz, identical cooperation from their colleagues. Now observe that the bargaining process above may yield an insider wage that exceeds the reservation wage by more than the ability-related marginal product differential, so that the outsiders are involuntarily unemployed. Nevertheless, the firms may have no incentive to replace the insiders. The reason is that the insiders and outsiders do *not* face identical external conditions of employment: the insiders receive cooperation whereas the outsiders (through no fault of their own) do not. Thus, the firms do not find it worthwhile to hire outsiders and consequently the unemployment persists.

By the same token, laid-off workers may be unable to retain their jobs by offering to work for lower wages. Specifically, suppose that there is a business downturn and that firms respond by laying off a number of employees. It can then be shown that it is in the best interest of the remaining employees to withdraw cooperation from the laid-off workers and thereby prevent underbidding.

Harassment activities can achieve a similar purpose. We observe that employees are free to decide how friendly or unfriendly they should be to fellow workers—activities whereby they can affect each other's disutility of work, but which firms usually cannot obtain complete, verifiable, and objective information about. Insiders can keep unemployed and laid-off workers from underbidding by creating the credible expectation that underbidders will be harassed. As a result, outsiders have a higher reservation wage than the insiders.

If the outsiders were able to avoid harassment, they would be willing and able to do the insiders' work for less than the insiders' wage. Yet they do not have this option. Their choice set, even allowing for their abilities, is less favorable than that of the insiders. Thus the unemployment may be regarded as involuntary, in much the same way as a person involuntarily relinquishes his wallet when a mugger asks him to "choose" between his wealth, and his health.

III. Effort and Labor Turnover

A third significant reason why firms might not comprehensively replace their high-wage insiders with low-wage outsiders is that the implied labor turnover would have an adverse effect on the morale of all their employees and consequently work effort and productivity would fall. As in some versions of the efficiency wage theories, we assume that firms have incomplete information on work effort and thus wages cannot be made dependent on it. Insiders know this and raise their wage above the level at which outsiders would be willing to work, but firms do not replace them since the associated productivity loss would dominate the reduction in labor cost.

To drive this point home in a simple way, suppose that future productivity is stochastically related to current work effort (due to lags in production or monitoring). Thus, firms cannot use current wages to reward workers for their current effort; at best, they can reward the stochastic output response to past effort. They can also use the turnover rate to stimulate effort by specifying a cut-off productivity, below which an employee is dismissed.

Let the firm's remuneration package consist of (a) a wage (which may be time rate and/or piece rate), and (b) the cut-off productivity. The firm can raise the labor turnover rate by raising its cut-off productivity. This reduces the expected future reward which each employee receives for current effort. It can be shown that the effort response depends on a substitution effect and an income effect (see our 1984a paper). By the former, effort falls: the employee works less hard since he is more likely to be fired and thus less likely to be compensated for his effort. The income effect raises effort: the employee works harder in order to avoid the possibility of being fired.

In this context, turnover has an adverse effect on effort if the substitution effect dominates the income effect. Whenever this is the case and insiders capture some of the economic rent associated with the turnover cost, then there may be involuntary unemployment in the following sense: Insiders and outsiders have the same abilities and differ only in terms of their competitive positions. If the outsiders could gain employment without affecting employees' effort incentives, they could perform the same job as the insiders—and do it for less than the insider wage. But since that option is closed to them, they may be considered involuntarily unemployed.

IV. Union Activity

Thus far our explanation of involuntary unemployment has not only avoided the presumption of government regulations, but also has made no reference to the activity of labor unions. However, the insider-outsider approach does suggest how unions may accentuate involuntary unemployment. It also provides several rationales for union activity and indicates how each can contribute to involuntary unemployment.

Assuming that unions are more responsive to the interests of their employed members than to the unemployed ones, there are many ways in which a union can help raise the wages of its insiders without reducing their chances of continued employment: (a) it may amplify the costs of hiring and firing (for example, severance pay, hiring and firing procedures); (b) it could increase the effectiveness and variety of cooperation and harassment activities; (c) it can augment insiders' bargaining power and thereby enable them to capture a greater share of the available rent from their jobs; (d) it can provide insiders with new rent-seeking tools: threats of strike and work-to-rule are the most prominent examples. (See our 1986, 1984b papers.) In this manner, the insider-

outsider approach offers an explanation how unions get their clout, and why employers choose to negotiate with unions rather than turn to nonunionized workers.

In short, the insider-outsider contributions described above may be seen as an attempt to rationalize simultaneously the existence of wage norms, involuntary unemployment, and the economic role of labor unions.

REFERENCES

Lindbeck, A. and Snower, D. J., "Involuntary Unemployment as an Insider-Outsider Dilemma," in W. Beckerman, ed., *Wage Rigidity, Employment, and Economic Policy*, London: Duckworth, forthcoming 1986.

______ **and** ______, (1984a) "Labour Turnover, Insider Morale, and Involuntary Unemployment," Seminar Paper No. 310, Institute for International Economic Studies, Stockholm, 1984.

______ **and** ______, (1984b) "Strikes, Lockouts, and Fiscal Policy," Seminar Paper No. 309, Stockholm: Institute for International Economic Studies, 1984.

______ **and** ______, (1985a) "Cooperation, Harassment, and Involuntary Unemployment," Seminar Paper No. 321, Institute for International Economic Studies, Stockholm, 1985.

______ **and** ______, (1985b) "Explanations to Unemployment," *Oxford Review of Economic Policy*, No. 2, 1985, *1*, 34–69.

Mitchell, D. J. B., "Explanations of Wage Inflexibilities: Institutions and Incentives," Working Paper No. 80, UCLA, 1985.

Okun, A., *Prices and Quantities*, Washington: The Brookings Institution, 1981.

Perry, G. L., "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1:1980, 207–41.

Solow, R., "Insiders and Outsiders in Wage Determination," *Scandinavian Journal of Economics*, No. 2, 1985, *87*.

Union Wage Rigidity: The Default Settings of Labor Law

By MICHAEL L. WACHTER*

Current discussions of wage norms begin with George Perry's analysis (1980). He argued that the rate of wage change appeared to shift in discrete steps. Although his study concerned aggregate wage adjustments, the union sector has long been identified as the primary source of wage rigidity and hence of wage norms.

The stylized explanation for the rigidity of any market price in the efficient contracting literature is the presence of high transaction costs. High transaction costs emanate from the internal rather than the external labor market. These costs make it inefficient to update wages continuously to changing market conditions. Alternatively stated, wage rigidity or norms is a Nash equilibrium for firms under ordinary circumstances (see Costas Azariadis, 1985). The equilibrium position is maintained until the transaction costs of making the change are less than the costs of maintaining the old regime. Once the regime changes, however, the new regime or equilibrium can be a discrete rather than a marginal change from the prior regime.

The regime in the unionized sector today is one of concession bargaining. Although concessions have been concentrated in those sectors that have experienced competition in a setting of deregulation and increased international trade, increased competition is more likely to be a consequence of earlier relative wage and cost changes than an exogenous cause of concessions today.[1] In fact, the common thread that binds together the industries that have exhibited concession bargaining is that they have emerged from a prior regime of significant and prolonged increases in union wage premiums. Indeed, as shown by Peter Linneman and myself, the increases in union wage premiums have caused a statistically significant and quantitatively large decrease in union employment. Concession bargains thus represent a shift in regimes as the parties attempt to deal with the effects of the prior regime of increasing premiums.

There is little doubt that the increase in union premiums was related to the supply shocks of the 1970's; that is, while nonunionized wages declined in response to these shocks, union real wages continued to increase. In this sense, the puzzle is why the unionized sector did not shift to lower wage norms during the 1970's and not the presence of concessions today.

The expansion of union wage premiums over the past decade cannot be explained by the traditional model of union-nonunion wage differentials. That model states that premiums remain steady over time unless changes occur in the underlying labor demand elasticities (i.e., the Hicks-Marshall conditions change) or the unions' tastes (reflecting the wages-employment tradeoff). Labor markets, however, have become more rather than less competitive in the 1970's as a consequence of deregulation and increasing international trade.

In disequilibrium, variation in the union wage premium can occur due to the fixed-contracting period. Indeed, union wage rigidity is typically explained by the existence of 3-year contracts that permit only incomplete wage adjustments during the contract period.

In this paper, the extent of union wage rigidity is shown to be related to contracting lags, but the lags are not identified with contract expiration and renegotiation dates. Rather, the lags and the resulting wage rigid-

*Professor of Economics, Law, and Management, University of Pennsylvania, Philadelphia, PA 19104. Costas Azariadis, David Hall, Clyde Summers, Lea Vandervelde, and Susan Wachter provided many helpful suggestions, and Rodrigo Quintanilla and Nancy Zurich provided valuable research assistance. The research was supported by the Institute for Law and Economics, University of Pennsylvania.

[1]See Peter Linneman and myself (1986) for a discussion of the endogeneity of increased international competition and, to a lesser extent, deregulation.

ity are closely related to the labor law governing collective bargaining. The union-employer relationship is, after all, regulated. Although the regulations do not fix wages, they do govern the adjustment process itself.

Internal labor markets, with their associated bilateral monopoly problem between a firm and its specifically trained workers, exist in the nonunion as well as in the union sectors. Given a perception of unequal bargaining power in internal labor markets, the law attempts to balance the scale by giving workers entitlements, the most important being the rights to bargain collectively and to impose costs on their employers if the bargaining breaks down.[2]

The law does not resolve the problem of high transaction costs resulting from bilateral monopoly. Rather it distributes entitlements differently from what would exist under the common law of contracts. Moreover, the entitlements are protected by property rules; that is, the parties must bargain with each other. This setting is quite different from typical commercial contracts or implicit labor contracts in nonunion firms. The result is an increase in transaction costs and hence in wage rigidity.

In Section I, the contracting process in unionized firms is analyzed in a law and economics, high-transaction-cost setting. In Section II, recent contracting developments are studied in terms of the parties' attempts to deal with the employment impact of high premiums.

I. The Contracting Process in Unionized Firms

The analysis of contracting lags in the unionized sector has focused on multiyear contracts with an explicit expiration point of 3 years. The 1973 to 1983 increase in union wage premiums, however, extended over a 10-year period. Hence, lags of 3 years are obviously too short. Substantial transaction costs must exist across contract expirations. In addition, these transaction costs must provide a setting not only for continuing, but also for sizable increases in union wages relative to nonunion wages.

A fundamental principle of labor law is that the parties are engaged in a long-term relationship and that the contract, although open to modification, runs with the relationship. Once a bargaining unit is certified, the firm must deal with the union as the agent for the workers. The law regulates this relationship by governing the procedure by which the parties must deal with each other.

In this context, contract renegotiations have a limited function. At contract expiration, the parties do not bargain de novo for a new contract. Rather, the parties negotiate over the contract that already exists. Moreover, the default setting is that the contract rolls over unless the terms are expressly changed by the parties. The burden to change the contract terms is thus on the initiator of the change.

As a consequence of the precedents that evolve from past contracts, the parties develop expectations of future performance. This is similar to that existing in nonunion contracts; the difference is the ability of the union to prevent the firm from acting unilaterally to defeat its expectations. Given the high-transaction-cost setting, pressing demands for major changes in the contract raises the possibility that one party will interpret it as an attempt by the other to seize the quasi rents of firm-specific investments, an act that might lead to reprisal.

An example of the contract rollover is that a firm cannot unilaterally implement contract changes, even after the contract expires, unless it first bargains to impasse. Even at impasse, the firm can only implement terms that have been expressly offered to the union. In addition, although the firm can hire new workers to replace striking workers, it must pay them the contract wage.

This contractual setting is obviously very different from that found in commercial contracts between buyers and sellers. In a typical commercial contract, the buyer (or seller) is free to terminate the relationship and find another seller when the existing contract expires. The buyer simply notifies the seller that the contract will not be renewed. In the collective bargaining contract, on the other

[2] See, for example, Oliver Williamson et al. (1975), Douglas Leslie (1984) and Richard Posner (1985).

hand, the buyer of labor cannot simply terminate the relationship when the contract expires and hire new suppliers of labor. The firm must deal with its current suppliers and their agent (the union). If the firm's dealings show "anti-union animus," its actions can be enjoined.

The above discussion describes the setting for employer-union contracting over the unanticipated supply-price shocks. The question is how do the parties adjust when the contract is unclear as to which party is liable to carry the burden of the shock?

Where a commercial contract does not clearly indicate which party is liable for particular market shocks, one party will believe that the other has breached the contract.[3] The aggrieved party has access to a court determination of fact as to whether breach has occurred; that is, a determination of which party is liable. If the contractor is liable, assessed damages are typically lump sum and thus long-run marginal costs are unaffected. Moreover, assessed damages rarely fully compensate for the effects of breach.

In labor law, on the other hand, the parties must bargain with each other over the liability of the supply-price shock burden. Neither party, acting unilaterally, can ask a court or the National Labor Relations Board for a determination of how the contract distributes the liability. Labor law regulation only deals with the enforcement of the negotiation process and not with the outcome of the negotiations.

Absent guidance from the court, but given the duty to bargain, the parties can be viewed as settling their own dispute by looking to the set of precedents guiding the collective bargaining parties in the past. This leads to an adjustment pattern that is autoregressive and where the short-run impact of new shocks is greatly attenuated. In this setting, continuing performance of union contracts with high premiums means ongoing payments of damages that affect the firm's marginal cost.

[3] Supply shocks caused similar problems in commercial contracts. For a discussion of the issue, see Paul Joskow (1976).

In summary, the legal setting between the firm and its unionized workers can be depicted as an open-ended, long-term, buyer-seller contracting. The parties must deal with each other to resolve a dispute. The high transaction costs are thus those of a bilateral monopoly where the entitlements of the parties are protected by property rules (i.e., the parties must bargain together) rather than liability rules (i.e., the court determines and assesses damages).

II. Recent Contracting Developments

To understand the adjustment mechanisms, the union and nonunion responses to shocks can be differentiated. Nonunion as well as union firms develop internal labor markets and implicitly contract over the long term with their specifically trained workers. There is, however, no agent for the workers with whom the nonunion firm must bargain. Hence, nonunion firms exercise more freedom in adjusting their wage bill in response to shocks. Empirically, these firms appear to have contracted with their workers over the past decade so as to reduce employment fluctuations at the cost of higher wage variance. Workers, in effect, accept the risk of fluctuations in real wages in order to reduce the risk of job loss resulting from the supply shocks.

Union contracts can be interpreted as working differently in that workers' expected future payments were based on real wage trajectories existing prior to the supply shocks (perhaps as a return for previous firm-specific investments by the workers). Thus real wages in the unionized sector continued to increase despite the supply-price shocks. In return for this stability of real wages, workers, in effect, accepted greater risk of job loss. Although much of this risk was assigned to junior workers, the current environment shows that considerable risk to senior workers remained.

The change in regimes from increasing premiums to concessions means that the costs of maintaining the old regime are greater than the transaction costs of making the change. Hence, unions must determine whether to attempt to maintain the high premiums that emerged during the 1970's

and early 1980's, or to accept the cost of continuing declines (or stability at low levels) of employment. A policy of maintaining the high wage premiums is one form of "end-game strategy" (Colin Lawrence and Robert Lawrence, 1985).[4] Given the historically high union premiums now existing, an end-game strategy of seizing the quasi rents may seem to unions to be the best of bad alternatives.

If the labor unions remain adamant about maintaining wage premiums of the current magnitude, firms have several options. A novel approach is the two-tier wage structure which allows tenured union workers to be compensated for past firm-specific investments at the preshock real wage trajectory and new workers (not "owed" back wages) to be paid based on the lower current trajectory.

A second option is to fight the union by taking a strike and, if that does not succeed, by using various legal options such as replacing strikers or asking for a new election to decertify the union. This solution represents a large intramarginal shift in equilibria, but when the costs of continuing to pay the old wage trajectory are large enough, this strategy may be adopted by the firm.

A third option, which is likely to be the most utilized, is to shift capital out of the market in which it is in a noncompetitive position. To date, research on union premiums has focused on manufacturing, where the nonunion sector is outside of the United States and capital is relatively immobile. Most of the U.S. union employment decline, however, is in sectors such as transportation, construction, and even retail trade, where relatively mobile capital has been shifted to nonunion U.S. firms. In terms of wage norms, the decision to exit the industry represents a discrete shift in the equilibrium position that takes place only after the cost of continued performance surmounts the hurdle posed by high transaction costs.

Not only do firms and unions adopt new strategies, but the legal setting itself may reinterpret precedents in response to pressures generated by sticky wage norms. In *UAW v. Milwaukee Spring*, a relevant case to the issue of capital mobility between union and nonunion sectors, the court approved an NLRB decision that permitted a firm to shift work from its unionized plant to its own nonunion subsidiary in the middle of a contract. The NLRB ruled that management reserved that right unless it was otherwise expressly bargained away in the contract.

To an extent, the change of this legal rule implicity reflects the buildup in the premium. Firms are better able to meet the legal requirements for outsourcing when premiums are large. If a firm shifted work to a nonunion plant absent large premiums or cost differentials, the shift would more likely be viewed as showing anti-union animus.

In summary, adjustments in collective bargaining take place but often in discrete steps. The parties' ability to make marginal adjustments is limited by high transaction costs and the associated parameters of labor law. Once the transaction cost hurdle is made, however, larger discrete changes can occur; hence the shift in norms from expanding premiums to concession bargaining.

[4]For a critique of the end-game model as applied to the expansion of the union wage premium, see my 1985 paper.

REFERENCES

Azariadis, Costas, "Rational Expectations Equilibria with Keynesian Properties," mimeo., University of Pennsylvania, November 1985.

Joskow, Paul L., "Commercial Impossibility, the Uranium Market and the Westinghouse Case," *Journal of Legal Studies*, March 1976, *6*, 143–74.

Lawrence, Colin and Lawrence, Robert Z., "Manufacturing Wage Dispersion: An End Game Interpretation, *Brookings Papers on Economic Activity*, 1:1985, 47–116.

Leslie, Douglas L., "Labor Bargaining Units," *Virginia Law Review*, April 1984, *70*, 353–418.

Linneman, Peter D. and Wachter, Michael L., "Union Wage Premiums and Employment," mimeo., University of Pennsylvania, 1986.

Mitchell, Daniel J. B., "Shifting Norms in Wage Determination," *Brookings Papers on Eco-*

Shifting Wage Norms and their Implications

By GEORGE L. PERRY*

At least since the early 1970's, it has been apparent that the cyclical variations in inflation summarized by the short-run Phillips curve are only one part of the inflation problem that confronts modern industrial economies. Another part is the relative persistence of an established rate of inflation. There is a good deal that we do not understand about this persistence. But I find the most useful way to model it is to start with the concept of a relatively stable wage norm, by which I mean a norm for the rate of wage increase. The model distinguishes sharply between the cycle and the trend in inflation, with the wage norm determining the trend. The variations in inflation of the typical business cycle take place around the existing wage norm and generate the empirical short-run Phillips curve. The wage norm itself is affected little if at all by the typical business cycle. Historically the wage norm has been shifted by prolonged departures from typical business cycles or by other extreme economic developments. Figuring out more precisely what it takes to shift wage norms, or what might keep them from shifting, is a central challenge for understanding inflation better.

Before turning to its implications, let me sketch the behavioral underpinnings of the wage norm model and the empirical evidence about wage norms. The norm rate of wage increase has no allocational significance and describes the trend of nominal wages independent of real aggregate demand or relative demand effects. In this respect, it is like the anticipated rate of inflation in many familiar models. Wages are not determined in an auction-like labor market that clears over any reasonable interval of time. Rather they are established by wage-setting firms with a profit-maximizing interest in their long-run relation with their employees, in some cases in a bargaining situation with unions. Under both the implicit and explicit contracts that thus dominate wage setting, keeping up with the norm is the neutral standard for firms.

An individual firm that raises wages in line with the norm neither improves nor worsens its relative position as an employer. A firm that wants to expand employment will, typically, offer a higher wage than would be required just to keep up with the norm. Relative wages and relative employment levels are thus codetermined in this process. When most firms want to expand employment, as in a cyclical upturn, the same behavior is part of the process producing the modest cyclical rise in inflation that we observe as the short-run Phillips curve. Thus the onset of cyclical inflation is not a sign that capital and labor resources are being overutilized. Nor is it a sign that inflation is on an accelerating path or even that wage norms are shifting up.

In analyzing U.S. postwar data, I have found the wage norm shifted up substantially by the end of the 1960's and down again, though not by as much, by the end of the 1980–82 recession (see my 1983 paper). The first episode was a period with a historic expansion that ended with several years of very low unemployment rates. The second was a recession of unusual length and severity that ended with the highest unemployment rates since the 1930's. There is also evidence of a small shift down in the wage norm after the weak economic performance of 1957–61, which featured two recessions with only an aborted recovery in between. I also found evidence for Germany, the United Kingdom, and Japan of upward shifts in wage norms in manufacturing industries after the 1960's and downward shifts in the early 1980's (see my 1986 paper). All these episodes suggest the kinds of extreme cyclical developments that have shifted wage norms in the postwar period.

*The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036. In preparing this paper, I benefited from discussions with Charles Schultze.

Charles Schultze (1986) has analyzed U.S. price inflation experience from 1871 to the present, leaving aside the two world wars and the Great Depression. These omitted events presumably altered wage norms, but in a way that would confuse rather than illuminate an analysis aimed at more normal periods. There were numerous ups and downs of the business cycles in the years before the depression. Indeed, cycles were more frequent, on average, than in the postwar years. Yet Schultze finds, strikingly, only one episode resembling a norm shift: late in the nineteenth century, the norm inflation rate went from negative to positive for reasons that elude my understanding. The absence of shifts over any other interval in this long period supports the idea that norms are little, if at all, affected by typical business cycles. Again excluding the two world wars and the Great Depression, Schultze also finds that the cyclical relation represented by the short-run Phillips curve has been remarkably stable over the long period he analyzes, once the few norm shifts I have just described are allowed for.

I now turn to some implications of the wage norm model and the empirical record I have just outlined. One implication has to do with empirical modelling of inflation. Inflation has usually been modeled by combining a short-run Phillips curve with a model of adaptive expectations in which current wage changes are related to past values of price inflation. In these models, the cyclical response of wages, as captured in the short-run Phillips curve, is gradually and continually modified by the inflation of the past. I have found (1980) empirically that allowing for nonlinear norm shifts after periods of extreme economic performance fits the data somewhat better than a stable adaptive expectations model. This finding cautions against the estimates provided by a simple linear model such as adaptive inflationary expectations, even though it does not tell us how to build endogenous norm shifts into empirical work because they occur too infrequently to allow for conventional statistical modeling of the process that gives rise to them.

If wage norms are indeed relatively stable under normal conditions but shift in a nonlinear way in response to extreme economic developments, adaptive expectations models will exaggerate the impact on wages of past inflation at most times and understate the movements in wages under extreme conditions. As a related point, models that mechanically relate wages to past inflation will not distinguish between the effects on wages from exogenous price shocks, such as the fuel price explosions of the 1970's or changes in the terms of trade, and the inflation or disinflation that accompanies extreme business cycle developments. Yet wage norms may in fact be responsive to the latter situation but not to the former type of price development.

The distinction between adaptive expectations and norm shifts also has important implications for how other noncyclical events affect inflation. As one example, a slowdown in the productivity trend will convert any given rate of wage increase to a faster rate of price inflation. An adaptive expectations model will feed the faster inflation rate back into wages and predict that higher unemployment is needed, permanently, to maintain the original rate of inflation. In the usual formulation of such models, with a natural unemployment rate and an elasticity of wages to lagged price inflation of 1.0, inflation will accelerate indefinitely at the original unemployment rate in response to a permanent slowdown in productivity growth.

Cross-country comparisons, as well as common sense, reject any such long-run connection between productivity trends, inflation, and unemployment. Wage norms differ across countries, in part reflecting differences in productivity trends. A change in a country's productivity trend would be expected, eventually, to change nominal wage norms or to raise the steady, nonaccelerating rate of price inflation. How near the response came to one or the other outcome would depend on behavioral response of firms and workers and, perhaps, depend also on the unemployment rate and other cyclical conditions that prevail. If widely recognized, a slower productivity trend might induce a prompt cut in the wage norm used by wage-setting firms

and induce workers to accept the slower real wage gains that a lower nominal wage norm would produce. Alternatively, competitive conditions might progressively squeeze profit margins for a time before leading either to a faster rate of price increase or to a lower wage norm. These are the kinds of questions about norm shifts for which we do not yet have reliable answers.

Turning now to stabilization issues, the wage norm idea has potential implications at three levels. First, how will inflation in industrial countries behave from here now that norms have shifted down? Second, in the present economic environment, what output-inflation scenarios would cause norms to shift up again? Third, and most conjectural, are there changes in the economic environment that could alter the behavior that has generated relatively stable norms and thus augment or diminish the stability of wage norms?

At the first level, the immediate message for short-term stabilization policies is that fear of inflation need not inhibit policymakers from pursuing economic expansion. The cyclical component of inflation described by short-run Phillips curves can be expected to appear. And exchange rate movements will add to price levels in some countries and hold prices down in others. But the component of inflation represented by the high wage norms that existed in the early 1980's has been sharply reduced. The wage norm model predicts it will not reemerge if economies now undergo a normal cyclical expansion from present levels of unemployment and operating rates.

At the second level, the message of the 1960's cautions against eventually pushing these expansions too far for too long. A major part of the economic trials and tribulations of the 1970's stems from the sporadic and costly efforts to reduce the high wage norms that were inherited from the overly strong expansions of the 1960's. On the basis of the experience, an extended period of overly tight markets should be avoided because it would risk ratcheting up wage norms again. However, the empirical record cannot tell us exactly how far expansions can be pushed before they threaten to shake the norm loose from its mooring.

At the third level of possible implications, the entire long cycle of the postwar period —which took most of the industrial world from the shambles of war, through a prolonged period of great prosperity, through high inflation rates and finally to prolonged and deep recessions—could represent a changed environment with consequences for the stability of wage norms. The outstanding feature of this new environment, particularly starting in the 1960's, was the commitment to high employment on the part of governments. It is conceivable that this commitment, as it succeeded in maintaining uprecedented prosperity, eventually made wage norms less stable than they had been in the past. That is, not only did wage norms ratchet up after the 1960's—a case of norms shifting predictably within a prevailing economic environment—but the stability of wage norms was lessened—a case of a changed economic environment altering wage-setting behavior. For now, this possibility is no more than a conjecture.

New classical models stress the importance of credible expected policies for determining the inflation and unemployment outcomes in the economy. One could relate them to the wage norm model either as models in which wage norms are determined by expected policies, or as models in which a credible change in policies alters wage-setting behavior in a way that increases or reduces the stability of wage norms. But, although such interpretations are possible, the new classical models have different behavioral underpinnings and different implications from the wage norm model.

As I interpret the new classical models, their central implication for now is that the deep recessions of the 1980's were necessary for establishing the commitment of policymakers to fight inflation. The belief is that economies will be less inflationary in the future because policymakers credibly demonstrated their willingness to impose deflationary pain this time. The deceptively simple prescription of the new classical models is to abandon employment goals in favor of poli-

cies directed exclusively at price stability. But if the argument is simply that imposing real deflationary pain in the last cycle will allow more prosperity pleasure in the next one before inflation reemerges, it is a prescription without a cure. It must imply that two cycles from now more pain will be needed to condition the economy away from the pleasure enjoyed in the next one.

Thus to be interesting the argument must imply more than this—specifically that policies aimed exclusively at price stability will also produce optimal levels of real activity. that may be regarded as an assumption of new classical models. But it is not presumed in the wage norm model. Nor is it the verdict of the data. The experience of the 1980's, particularly in Europe, disproves the idea that steady anti-inflation policies will not depress output and employment for long and that such policies are thus both necessary and sufficient for achieving prosperity and price stability together. Indeed the whole historical record shows that employment goals are not attained automatically through price flexibility. They must be regarded as objects of policy if they are to be achieved. In the context of the wage norm model, that implies finding policies and policy dosages that will keep wage norms low and stable at the same time economies are allowed to operate at efficiently high levels of employment.

That may prove doable if the relative stability of wage norms is still intact. In this case policy would mainly have to avoid repeating the excessively strong demands and tight markets of the 1960's. Since economic advisers recognized the need for restraint by the mid-1960's, this is hardly a counsel of unattainable perfection or a prescription that can be written only with hindsight. However, if it turns out that a renewed commitment to employment goals itself alters wage-setting behavior in a way that destabilizes wage norms, economists and policymakers may have to explore innovative ways to make prosperity compatible with acceptable inflation performance. Tax incentives for maintaining a low norm in wage-setting may be one avenue to explore. Another may be profit sharing or the more sophisticated compensation schemes developed by Martin Weitzman (1985) which would alter wage setting behavior in a fundamental way.

REFERENCES

Perry, George L., "What Have We Learned About Disinflation?," *Brookings Papers on Economic Activity*, 2:1983, 587–602.

______, "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1:1980, 207–41.

______, "Policy Lessons from the Post-war Period," *Wage Rigidity, Employment, and Economic Policy*, Oxford: Oxford University Press, forthcoming 1986.

Schultze, Charles L., *Inflation in the United States and Europe: An Historical Comparison*, Washington: The Brookings Institution, forthcoming 1986.

Weitzman, Martin L., "The Simple Macroeconomics of Profit Sharing," *American Economic Review*, December 1985, *75*, 937–53.

Union vs. Nonunion Wage Norm Shifts

By DANIEL J. B. MITCHELL*

Empirical investigations of wage determination have often produced autocorrelated residuals from time-series wage equations. Runs of over- or underprediction have usually been regarded as weaknesses in specification to be corrected or explained away. In 1980, however, George Perry suggested that such runs represent an important, if neglected, characteristic of American wage setting. He argued that "norms" of wage change develop in the labor market. These norms, according to Perry, change discretely; there are periods of more or less wage "pushiness."

Aggregate wage indexes can be influenced, even if norm shifts are not fully reflected everywhere, providing those sectors that are affected have sufficient weight in the indexes. An obvious division in the labor market is between the union and nonunion sectors. There is reason to believe that while there has been a (downward) shift in wage norms recently, the impact has been concentrated in the union sector (see my 1985 article). Indeed, the union sector is probably inherently more prone to norm shifts than the nonunion.

I. Union Sector Developments

Table 1 is supportive of this proposition. The table shows wage trends in major union agreements and in an index of selected nonunion wages over various subperiods covering the years 1961–84. The union wage settlements are those involving large numbers of workers and are often considered to be trend setters for other union negotiations. The industries selected for the nonunion series are those where there is very little unionization and union activity; they are "purely" nonunion in the sense that there is little likelihood of union spillover or threat effects in their wage decisions.

TABLE 1—TRENDS IN UNION AND NONUNION WAGES, 1961–84

Period	Median Adjustments in Major Union Agreements	Adjustments in Earnings of Selected Nonunion Industries
1961–64	2.8	3.3
1965–69	4.4	5.6
1970–75	7.8	5.6
1976–79	8.4	7.0
1980–82	9.3	7.3
1983–84	4.2	4.8

Note: Nonunion series is percent changes in annual average hourly earnings in SIC 533 (variety stores), SIC 56 (apparel stores), SIC 57 (furniture stores), and SIC 60 (banking). Figures for the four industries are averaged using 1979 nonsupervisory employment weights. The figures are simple averages of yearly data for each period, shown as annualized percent changes.

Union wages exhibit more variation than nonunion over the period covered. For example, the nonunion wage series shows less acceleration during the inflationary 1970's than the union series. In the most recent subperiod shown—years when the union wage concession movement was in full swing—the nonunion wage index shows less deceleration from the past and outpaces the union index.

First-year union wage freezes and cuts began to appear in significant numbers in 1981. During the first 9 months of 1985, 3 years after the 1982 recession trough, freezes and cuts still represented over one-fifth of new union settlements. Various devices have been introduced in the union sector to accommodate a downward norm shift. These include introduction of "two-tier" pay plans (which provide lower wage schedules for new hires), and the substitution of lump sum bonuses for increases in base wage rates. While elimination of cost-of-living adjust-

*Director, Institute of Industrial Relations, UCLA, Los Angeles, CA 90024.

ment clauses (COLAs) has not been especially common, the restriction of COLA formulas to provide less money has become widespread.

One characteristic of union wage setting is the use of long-duration contracts. So far there has been little evidence of a substantial shift to shorter contracts during the concession period. There has been a limited move toward profit sharing in union contracts during the concession era. Both developments represent management preferences; management dislikes like short contracts because of the strike-related uncertainty of frequent renegotiations. But management has evinced an interest in sharing product-market risk with workers via profit sharing.

II. Historical Evidence

Economists generally view the labor market as dominated by an impersonal, invisible hand. Yet the norm concept suggests that shifts in the balance between organized labor and management—sometimes supported by the external legal and political environment—play an important role in union wage outcomes. Wage norms have shifted down after episodes in which management comes to feel that unions have gone "too" far in pressing their claims. Union-nonunion wage differentials are one index—but not the exclusive index—of such episodes. Three periods of management counterreaction are reported on Table 2. In the earliest case, management's reaction was mainly a drive for the nonunion alternative. The second case—while not devoid of such goals—was largely characterized by a management attempt to hold back labor costs in the context of a bargaining relationship. Finally, the most recent episode has featured strong management thrusts in both areas: holding down wage increases and seeking nonunion operations.

During World War I, union membership rapidly increased under protective federal policies aimed at avoiding industrial unrest. Management reacted with a considerable effort to professionalize personnel policy in order to avoid unionization and reduce quits. Despite membership gains, union real wages initially slipped in the face of accelerating inflation. But union resistance to nominal wage cuts led to real wage gains in the immediate postwar deflation and a large jump in the union-nonunion wage gap. Management responded in the early 1920's with an "open shop" drive. The primary goal was to avoid unions altogether rather than close the gap. Company unions were created and unions were pushed from many major industries. The legal and political system was no longer supportive of unionization; quite the contrary. Faced with adversity, however, unions emphasized cooperation with management in certain prominent, well-publicized situations. As in the 1980's, cooperative themes in some industries coexisted with a management drive in others to remain or become nonunion. Strike activity declined.

Perry identifies the early 1960's as a period of a downward wage norm shift. In terms of the political climate, there was a marked difference between the 1920's and 1960's. But, there were also some parallels. Prior to

TABLE 2—WAGE TRENDS AND RELATED INDICATORS IN THREE PERIODS

Period	Ratio: Union-to-Average Wage, Percent Change	Annual Strike Index (First sub-period = 100)	Change in Union Membership (millions)
1917–20	1.5	100	+2.3
1921–22	25.1	47	−1.0
1923–26	4.7	35	−.5
1956–58	5.4	100	+.2
1959–60	.3	78	0.0
1961–64	−1.7	70	−.2
1976–79	2.0	100	+.2
1980–82	2.7	57	−2.8
1982–84	−1.1	32	−.4

Note: The union-to-average wage series refers to manufacturing for 1917–26 and to the private, nonfarm economy for later periods. For 1917–26, the Douglas union wage series has been divided by BLS average hourly earnings. For 1956–64, estimates based on BLS data for major union contracts have been divided by average hourly earnings. For 1976–84, data were drawn from the BLS Employment Cost Index for wages and salaries. The strike index refers to all reported disputes for 1917–26 and for major disputes (1,000 or more workers) for 1956–64 and 1976–84. Union membership changes are drawn from the Wolman series for 1917–26, from BLS membership surveys for 1956–64, and from a combination of data from the *Current Population Survey* and the Bureau of National Affairs for 1976–84. Details are available from the author.

the norm downshift, union-nonunion wage differentials widened. The Landrum-Griffin Act, which regulates union conduct, was adopted over union objections in 1959. National Labor Relations Board (NLRB) data suggest that management became more willing to test the limits of the legal system at about the time the wage norm began to shift. Management also evidenced a growing concern to regain control of the workplace through reduction in restrictive workrules, for example, in steel and railroads.

The beginnings of the relative slippage in unionization of the workforce were sufficiently marked by the early 1960's to spark an academic debate on the future of unionism. Concern about job loss led to numerous conferences on "automation," a foreshadowing of the "robotics" discussions of the 1980's. Foreign competition began to intrude in some industries, notably steel. Well-publicized cooperative experiments were undertaken, and strike activity declined. Federal wage guideposts reinforced the idea of wage moderation.

Demand pressure eventually overcame these tendencies during the Vietnam War era. Such pressures could have the same effect in the late 1980's, if they arise again. However, the legal and political environment in the 1980's is more adverse to unions in the current period than it was in the 1960's. The recent misfortunes of the labor movement have become such standard media fare that there is little need to cite them here. Suffice it to say that the administration of basic labor law has shifted substantially under the Reagan Administration. Mondale's electoral debacle had a demoralizing effect on AFL-CIO leaders. Alternatives to conventional bargaining as the major function of unions and to use of the NLRB framework are now openly debated in union circles.

As in previous periods, especially the 1920's, management attitudes toward unions have been partly conditioned by earlier labor relations developments. In the late 1960's, there was a wave of strikes and rank-and-file rejections of tentative settlements. This wave was followed in the 1970's by a widening of union-nonunion wage differentials. Management's reaction to this previous militancy is now apparent.

Researchers have recently found evidence of changes in management strategy in the late 1960's, that is, in the period when wage norms were shifting up, toward more concentrated union avoidance (Thomas Kochan et al., 1985). In some companies, these changes took the form of increased concern for "human resources." Substitutes for unions in the form of improved communications with employees were sought. It is hard to resist the parallels between the employee representation (company union) schemes of the 1920's and latter-day "quality circles"! In other cases, more overt actions were taken, ranging from citing new plants in nonunion areas to dismissals of union organizers. Management's efforts in the post-1979 economic slump were expanded to encompass concession bargaining as well as union avoidance.

III. Union Membership Losses

During the 1970's, an underlying erosion of unionization was masked by economic expansion. Growth in public sector unionization also tended to compensate for private sector slippage. However, in the early 1980's, absolute membership losses became substantial. Unions represented about 2 million fewer workers under major agreements in the private sector in 1984 than they did in 1979. The usual explanation given is a decline of employment in "smokestack" industries. However, using a 41-industry breakdown of nonsupervisory employment trends, I find that only about one-fourth of this drop was explicable by employment shifts (see my earlier paper). Most of it stemmed from declining union representation *within* industries.

Given membership trends, the union wage norm shift of the 1980's must be viewed as part of a larger phenomenon with potentially more lasting effects than the 1960's episode. Indeed, the most lasting effect of the current period may be greater weight in total wage setting of nonunion employers. In that regard, the 1980's more closely resemble the 1920's than the early 1960's.

Despite the downward trend, the union sector is still large enough to influence the major wage indexes. That is why aggregate

wage indexes—especially hourly earnings—have shown surprisingly low rates of wage inflation. (Hourly earnings data data exclude nonsupervisory workers who are largely nonunion.) Union workers have higher average pay levels than nonunion and tend to work more hours. Their payroll weight is thus higher than their employment weight. While the frequency of wage freezes and cuts will undoubtedly decline, and while some predict that management will eventually overreach itself and produce a pro-union backlash, neither of these effects is likely to lead to a substantial upward shift in union wage norms in the near term.

IV. Nonunion Norm Shifts?

There are reasons to doubt that the norm concept is as applicable to nonunion wage setting as it is to union. Empirical research indicates that nonunion wages are more responsive to short-run demand fluctuations than union. During the economic slump of the early 1980's, survey data suggest that nonunion managers quickly revised downward their planned pay adjustments as demand for labor fell.

For example, Hewitt Associates' surveys indicated that wage increases planned for 1982 were revised downward from a projected 9.0–9.3 percent as of the summer of 1981 to an estimated 7.6–7.9 percent by mid-1982. Even this midstream estimate from the 1982 survey is higher than the actual 6.4 percent reported by the BLS for white-collar (largely nonunion) employees. Thus, the nonunion managers apparently made still further downward revisions during 1982. To the extent that nonunion wage freezes and cuts were used, they tended to be of short duration (less than a year).

Nonunion wages are not set in long-term contracts. And the strategic uncertainties of bargaining are not present in the nonunion sector. Long horizons of union contracts require the negotiators to look for estimates of future general wage trends, a requirement conducive to a wage norm mechanism. In addition, the parties must estimate the other sides' willingness to inflict costs in the event of an impasse in the union sector.

Since it is costly to determine this willingness experimentally, one side or the other may accumulate unexploited bargaining power. Thus, management may have had more bargaining clout than it realized in the 1970's, a period in which the union-nonunion wage differential was allowed to widen. Once a few managements were forced to discover their power by the economic slump of the early 1980's, other managements were induced to test their own ability to extract concessions. Thus, concessions spilled out into industries such as supermarkets where it is hard to tell tales of deregulation, special cyclical sensitivity, or imports. Again, there is no counterpart to such phenomena in the nonunion sector.

Even if nonunion wage setting is not norm prone, it is sensitive to demand. Nonunion wage setting will reflect macro policy. So far, the monetary authorities have elected a policy of maintaining a loose labor market rather than risk renewed inflation. As long as that policy continues, the nonunion sector is unlikely to become a source of cost pressures.

REFERENCES

Kochan, Thomas A., McKersie, Robert B. and Katz, Harry C., "U.S. Industrial Relations in Transition: A Summary Report" in Barbara D. Dennis, ed., *Proceedings of the Thirty-Seventh Annual Meeting*, December 28–30, 1984, Madison: Industrial Relations Research Association, 1985, 261–76.

Mitchell, Daniel J. B., "Shifting Norms in Wage Determination," *Brookings Papers on Economic Activity*, 2:1985, 575–99.

Perry, George L., "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1:1980, 207–41.

Hewitt Associates, *Compensation Exchange*, various issues, 1981–83.

Illegal Aliens: A Preliminary Report on an Employee-Employer Survey

By BARRY R. CHISWICK*

Although most observers agree that the number of illegal aliens in the United States has grown sharply over the past two decades, exact numbers are not available. Estimates suggest that they may account for as much as 2 to 4 percent of the U.S. labor force. Surely the presence of such a large and growing component in the labor market must have far-reaching effects.

This paper condenses parts of my larger study of the illegal alien labor market (1985). Section I outlines several sets of alternative hypotheses regarding the adjustment, role, and impact of illegal aliens in the labor market and discusses the data appropriate for testing these hypotheses. A new data set is discussed (Section II) that includes a matched sample of illegal aliens and their employers, and a parallel sample of employers randomly selected from industry directories. Space constraints preclude providing more than a taste of the richness of the analyses that can be done with these data. Two examples are provided. The matched employee-employer data are used to analyze the wages of illegal aliens (Section III) and the two employer samples are used to study employer differences in on-the-job training (Section IV).

I. Alternative Hypotheses and Data Requirements

It is difficult to separate myth from reality in the discussion of illegal aliens in the labor market. Although it is generally agreed that illegal aliens tend to be low-skilled workers compared to legal immigrants and the native population, and the available data support this view, there is no consensus on other issues.

To some, illegal aliens take jobs that would otherwise be held by persons with a legal right to work in this country, with direct competition in the labor market between illegal aliens and the low-skilled native workers for whom they are close substitutes in production. This competition lowers the earnings and employment opportunities of low-skilled natives. To others, however, the labor market appears segmented, illegal aliens being a noncompeting group with regards to U.S. workers. According to this view, illegal aliens are employed in jobs that, in their absence, would not otherwise exist because the jobs would not be done (for example, the grass would not be cut or would be cut less often), the products would be produced using different techniques (for example, the introduction of new crop-picking machinery), or the products would be imported (for example, the flight of the garment industries to *LDCs*). If they are right, illegal aliens have no depressing effect on the earnings and employment opportunities of low-skilled native workers. A related set of hypotheses refers to the characteristics of the employers of illegal aliens. Are they underground ethnic-enclave employers in a "secondary labor market"? Or are they a microcosm of establishments in the same region and industry who do not employ illegal aliens?

Some view illegal aliens as exclusively low-skilled, low-wage workers whose illegal status results in their being trapped in dead-end career paths and in ethnic-enclave employment. As such, they are subject to employer "exploitation." Others, however, suggest that illegal aliens are absorbed into

†*Discussants*: Jean Baldwin Grossman, Mathematica Policy Research, Inc.; Frank de Leeuw, U.S. Department of Commerce; Michael Piore, Massachusetts Institute of Technology.

*Department of Economics and Survey Research Laboratory, University of Illinois at Chicago, Chicago, IL 60680.

the U.S. labor market in a manner similar to legal immigrants. That is, although their investments in U.S. on-the-job training may vary with the extent to which they view themselves as attached to the U.S. labor market, illegal aliens experience improvements over time in earnings, employment, and occupational status as do other U.S. workers.

Competing hypotheses can persist for a long period of time when there are no reliable data for choosing among them. There is a virtual absence of systematic and reliable data on the labor market activities of illegal aliens. In part, this is because illegal aliens have an incentive to avoid revealing their status to an interviewer or on a questionnaire. In addition, the analysis of illegal alien labor market behavior and impacts requires survey data on the characteristics of the employer that may not be known to a typical worker, particularly a recent immigrant. These characteristics include the wage structure of the firm, percentage of employees unionized, training opportunities, hiring policies and practices, and the racial/ethnic composition of the workforce.

Many policy questions and research hypotheses regarding illegal aliens cannot be addressed solely by examining data on the aliens and their employers. The extent to which employers of illegal aliens offer on-the-job training, are part of the underground economy, or produce different goods and services than other establishments (or the same goods and services but in a different manner) cannot be meaningfully addressed without a benchmark. The benchmark may be employers that are known not to have employed an illegal alien or employers that are randomly selected. Although the preference would be for the former, there is no way of determining that an establishment does not and has not employed illegal aliens. Thus, randomly selected employers may be the most meaningful benchmark.

II. The Survey of Illegal Aliens and their Employers

A survey was designed specifically to study the illegal alien labor market. It generated data on the illegal alien's demographic and labor market characteristics derived from the alien, as well as data on the employer and the workplace derived from the employer. The sample of illegal alien employers has been augmented by a random sample of employers. The method for obtaining these data, combining administrative records on the illegal aliens with a survey of their employers and another of randomly selected employers, was shown to be both cost-effective and statistically reliable.

Whenever an illegal alien is apprehended by the Immigration and Naturalization Service (INS), a Record of Deportable Alien, referred to by its form number, I–213, is completed. The form includes questions on the alien's demographic characteristics, nationality, immigrant status, and labor market characteristics. The alien's identification of the employer was used to obtain a sample of employers of illegal aliens. These employers were then interviewed about the characteristics of the establishment and its employees. Combining data on the alien's characteristics and the employer's characteristics generated a unique employee-employer matched data file.

The employers identified by apprehended illegal aliens can be fruitfully compared to employers randomly selected from lists of establishments as long as (a) not all establishments employ illegal aliens or (b) establishments employing more illegal aliens, other things the same, are more likely to be identified in the I–213 files. Based on these assumptions, the employer sample consists of two components, one set of employers identified by apprehended illegal aliens and the other set randomly selected from other lists of establishments.

To carry out this study, data were recorded from a stratified random sample of nearly 300 I–213 forms for employed male illegal aliens apprehended in 1983 by the INS Chicago District Office. The list of employers identified by these illegal aliens was combined with over 300 employers randomly selected from industry directories. The establishments were interviewed during early 1984. Neither the interviewers nor the establishments knew the sources for the sample or

that the research interest was the illegal alien labor market.

Although it had been expected that employers of illegal aliens would be much more hesitant about participating in a survey than randomly selected employers, the very small difference in the interview completion rates and in item nonresponse suggests that this was not the case. There were 406 completed (nonpartial) interviews, 15 partial interviews, and 76 refusals among the establishments. The interview completion rate may be defined conservatively as the number of completed interviews (406) divided by the total number of cases not deemed to be ineligible (524). Calculating the conservative completion rate, 77 percent of the establishments were completed interviews, 76 percent in the INS sample, and 79 percent in the general sample.

TABLE 1—REGRESSION ANALYSIS OF HOURLY WAGES FOR ILLEGAL ALIENS, 1983

Variables	Coefficient	t-Statistic
Alien's Characteristics		
YREXP (years)	0.019	2.21
YREXPSQ	−0.0003	−1.69
TIMEUS (years)	0.041	2.37
TIMEUSSQ	−0.002	−1.15
STUDENT	−0.251	−1.70
VISITOR	0.075	1.35
Employer's Characteristics		
UNION	13.835	2.70
HSWAGE ($)	2.417	2.18
MSTCOMSK	0.189	2.80
REST	−0.157	−2.77
SERVICE	−0.069	−1.42
R^2	0.485	
Adj R^2	0.449	
Number of Observations	170	

Note: Dependent variable is the natural logarithm of the hourly wage. See text for a description of the variables.
Source: My 1985 study, Tables 4–5.

III. The Matched Employee-Employer Data: Analysis of Wages

In this section the unique feature of the matched employee-employer data is used to analyze the wages of illegal aliens. This may be the first microdata regression analysis of wages that uses establishment characteristics generated by an employer survey. The regression analysis is based on the standard human capital earnings function modified for the analysis of immigrant adjustment.

In the absence of data on years of schooling, "experience" is measured by the number of years since age 15 (*YREXP*), rather than the usual measure, the number of years since leaving school. The measure of duration of U.S. residence (*TIMEUS*) is the number of years (including fractions) between the date of last illegal entry and the date of action (i.e., when the I–213 form was completed). Dichotomous variables are created for marital status (*NOTMAR* = 1 if not married), status at entry (*STUDENT* = 1 if student visa, *VISITOR* = 1 if visitor or other visa, with "entry without inspection," or *EWI* as the omitted category), and country of origin (Mexico, Europe/Canada, other Latin American, Asia, and other countries).

Several measures of the characteristics of the employer and the workplace are added to the regression analysis. These variables include the size of the establishment (*SIZE* = number of employees), the degree of unionization (*UNION* = the proportion unionized among nonsupervisory workers), the wage scale (*HSWAGE* = the average wage paid to a 25-year-old high school graduate who worked for the firm for 2 years), and recent immigrants (*IMMIGHIRE* = the proportion of those hired in the past year who immigrated within the past 5 years). Dichotomous variables are created for skill level in the establishment (*MSTCOMSK* = 1 if the most common male nonsupervisory job is a professional, technical, or skilled production occupation), type of ownership (*SUBBR* = 1 for a branch or subsidiary of a larger firm), and the industry (restaurant (*REST*) or other service (*SERVICE*), with manufacturing as the benchmark).

The regression analysis of the hourly wage rate (mean $4.60) is reported in Table 1 after the regression was recomputed without several of the statistically insignificant variables. Holding constant their own demographic characteristics, the wages of the illegal aliens are significantly positively related to the degree of unionization (*UNION*), the wage scale (*HSWAGE*), and whether the most common male nonsupervisory job is skilled

(*MSTCOMSK*). Going from a 0 to 100 percent unionization, for example, raises the hourly wage of an illegal alien, other things the same, by about 14 percent. This is on the same order of magnitude as the effects of unions on wages in industry and general population studies.

Some of the positive effects on wages of additional U.S. labor market experience arises from illegal aliens obtaining employment in higher-wage establishments (larger values for *HSWAGE*, *UNION*, and *MSTCOMSK*) with a longer period of residence (see my 1985 study, ch. 4). The partial effect on wages of U.S. labor market experience is therefore lower when the establishment's characteristics are held constant. Yet, total experience (*YREXP*) and U.S. experience (*TIMEUS*) are still highly significant variables ever after controling for employer characteristics (Table 1). Hourly wages are lower by over 20 percent for those who entered the United States under a student visa (*STUDENT*) compared to those who entered without inspections, perhaps because many of them are still enrolled in school. Other things the same, marital status and country of origin have no significant effect on the hourly wage and they are not included in the equation reported in Table 1.

The wages of illegal aliens are apparently not related to the size or type of ownership of the establishment or the extent to which recent immigrants were among those hired in the past year. The latter finding suggests that the ethnic enclave patterns of employment found among illegal aliens do not depress their wages, other things the same. There is no significant difference in wages between jobs in service establishments (*SERVICE*) and in manufacturing. Wages of illegal aliens are significantly lower in restaurants (*REST*) than in manufacturing. However, part of this 16 percent differential may be compensating for unmeasured dimensions of job-related income, such as unreported tips and income in kind in the form of free meals.

IV. The Employer Samples—Analysis of On-the-Job Training

Here I focus on the analysis of on-the-job training provided by the employer. The extent of on-the-job training received by illegal aliens in the United States is important for understanding their labor market adjustment, their degree of attachment to the U.S. labor market, and the types of jobs they hold. The survey included questions on the number of business days of training in the establishment that a newly hired worker generally requires to learn to do well the most common male nonsupervisory job. The questions were asked for workers with no prior experience in the job (*TRNOPRIOR*) and for those with prior experience (*TRYESPRIOR*).

The incentives for worker investments in on-the-job training that are specific to a firm, an industry or a country are related to the degree of permanence the worker attaches to employment in that firm, industry or country (see my 1984 article). For a variety of reasons, illegal aliens would anticipate a less permanent attachment to a particular firm or industry in the United States, or even to working in the U.S. labor market, than would otherwise similarly situated legal resident aliens or native-born workers. Illegal aliens would therefore prefer jobs whose training, if any, is more general and less specific. Thus, for example, compared to legal immigrants or the native born, training that is general to the U.S. labor market rather than specific to a particular firm would be relatively more attractive for illegal aliens. Even training that is general to the firm but specific to the United States would be less attractive for illegal aliens. Employers would similarly view illegal aliens as relatively less attractive workers for jobs that involve larger amounts of employer-financed firm-specific training.

It is therefore hypothesized that establishments in which the most common male nonsupervisory job involves lengthy or costly on-the-job training would be less likely to employ illegal aliens. The negative relationship between training and employment of illegal aliens is hypothesized to be stronger for firm-specific training than for general training. The number of days of training received by newly hired workers without prior experience in the job (*TRNOPRIOR*) would have relatively more general components than would the training received by newly hired workers with prior experience (*TRYES-*

PRIOR), so a stronger negative relation is hypothesized for the latter variable.

The statistical test is performed by regressing the dichotomous variable for whether the employer was identified by an apprehended illegal alien ($ILLEMP = 1$ if in the INS sample) on the days of training reported by the establishment, controlling for the total number of employees and the occupational level of these workers. The analysis shows that male nonsupervisory jobs that require a longer training period are significantly less likely to be in the INS sample, even after controlling for establishment size and the skill level of the job.[1] The negative relationship is stronger when relatively more of the training is firm specific (*TRYESPRIOR*).

Thus, the data are consistent with the hypothesis that male illegal aliens work for employers that provide less on-the-job training, particularly firm-specific on-the-job training. Yet, this does not lock illegal aliens into dead-end career paths. They do experience considerable job mobility and improvements in their employment opportunities with a longer duration of residence, in part due to general on-the-job training and in part due to acquiring skills relevant for the labor market merely by living and working in the United States.

[1]The means, standard deviations (shown in brackets), regression coefficients, and *t*-ratios (shown in parentheses) are (the source is my 1985 study, ch. 5):

Variable	Mean	Regression Coefficient[a]	Regression Coefficient[b]
TRNOPRIOR[c]	155.8	−0.00028	−0.00013
($N = 365$)	[258.3]	(−2.95)	(−1.28)
TRYESPRIOR[c]	45.3	−0.00089	−0.00054
($N = 386$)	[97.2]	(−3.62)	(−2.23)

[a]Dependent variable is *ILLEMP*, controlling for *SIZE* and it square.
[b]Dependent variable is *ILLEMP*, controlling for *SIZE* and its square, and the occupation of the most common male nonsupervisory job (seven categories).
[c]Separate regressions are estimated.

V. Summary and Conclusions

This paper reports on the development and preliminary analysis of a unique data file that includes matched employee-employer data for a sample of illegal aliens, and parallel data on the establishment and the workforce for employers randomly selected from industry directories. It is shown how these data can be used to address substantive issues regarding the labor market adjustment and impact of illegal aliens through two examples, analyses of wages and on-the-job training.

In addition to the myriad of specific conclusions that have been developed from the methodological and data analyses, two general conclusions emerge from this project. One is that the methodology is successful for investigating the labor market adjustment and impact of illegal aliens. More generally, it demonstrates the feasibility of developing employee-employer matched data files for other demographic groups or the labor force as a whole. The other is that the illegal alien labor market appears to be well-functioning; that is, it is competitive, fluid, and flexible, and provides opportunities for economic advancement and job mobility even for low-skilled foreign-born workers in this country illegally.

REFERENCES

Chiswick, Barry R., "Illegal Aliens in the United States Labor Market: Analysis of Occupational Attainment and Earnings," *International Migration Review*, Fall 1984, *18*, 714–33.

______, "The Employment and Employers of Illegal Aliens: The Survey and Analysis of Data," mimeo., Report prepared for the Sloan Foundation, University of Illinois at Chicago, July 1985.

Illegal Immigration

By WILFRED J. ETHIER*

Illegal immigration is the major international economic issue facing the United States. An idiotic initiative towards protection might well change this, but our trade problems are of our own making. Illegal immigration thrusts itself upon us, like it or not. The topic deserves formal treatment by economists, and this paper sketches out early steps toward that end.[1]

Although my treatment has been strongly influenced by American issues, it is increasingly relevant to a broader Atlantic perspective: European problems now look more and more like American problems, and reforms now discussed in the United States sound more and more like European practice.

Let us start with a very simple framework and see how far it can take us. Aggregate all output into a single good and ignore international trade. Illegal immigrants I are a perfect substitute in production for a group L of native workers (including legal immigrants), and an imperfect substitute for skilled labor S, inelastically supplied ($U = I + L$ denotes unskilled labor). Output is determined by a neoclassical production function $Sf(u)$, where u denotes unskilled workers employed per skilled worker.

Immigration determines I, so consider the migration decision. Suppose that potential migrants can earn a (fixed) wage equivalent w^* if they remain at home.[2] If they attempt to migrate, they face the probability g of being caught and denied entry, incurring a cost k. If they succeed, they face the conditional probability α of obtaining legal status (either fraudulently or through legal means, such as amnesty), in which case they earn the expected wage w_L of L workers. Otherwise they earn the wage w_I accorded members of I. The expected reward for attempting entry is therefore $g(w^* - k) + (1-g)[\alpha w_L + (1-\alpha)w_I]$. I assume that attempted migration proceeds until this expected reward equals the reward to staying put: w^*. This equality can be written:

$$(1) \quad \alpha w_L + (1-\alpha)w_I \equiv w = w^* + kg/(1-g) \equiv w(E).$$

where E denotes the total resources allocated to interdiction at the border: its magnitude influences the size of g.

The basic American policy tool has been border interdiction. What does this imply? First, if interdiction is the *only* enforcement policy, successful immigrants enter the same labor pool as legal workers, so that $w = w_L = w_I$. Then, from (1), E, through g, determines this common w. If domestic labor markets clear, w determines u, which in turn equals $(I + L)/S$. Thus successful immigration I is determined. Also an intensification of enforcement (an increase in E and therefore in g), will clearly raise w and thereby lower I and the wage paid to S while making unskilled workers better off.[3] Suppose on the other hand that the unskilled labor market has a downwardly rigid wage w_U. Firms hire workers by a random draw from the pool $U = I + L$. Then equation (1) again gives w,

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104–6297.

[1]A more detailed formal analysis, along the same lines as the present paper but with a different emphasis, may be found in my earlier article (1986).

[2]I take domestic output as numeraire.

[3]A complete description of the effects of an increase in E should also take into account who pays for the increased enforcement, a point related to the controversial question of whether illegal immigrants on balance contribute more to the government budget than they draw from it. If the tax burden falls entirely upon members of S, the distributional consequences mentioned in the text are simply accentuated. If, instead, unskilled labor forms part of the tax base, the effects of interdiction on the volume I of immigration and upon the incomes of skilled workers could become ambiguous. For more on this point, see my article.

which must equal ew_U, if e denotes the employment rate of unskilled workers. Also $f'(u)=w_U$, so that $u=eU/S$ then determines I. Clearly an intensification of interdiction will raise the employment rate e, lower I, and leave members of S unaffected (unless they pay for the higher E).

But why should a nation have an immigration policy? One possible answer is to exploit market power with respect to mobile labor, that is, to maximize national income. A second objective could be the internal distribution of income between skilled and unskilled workers. Finally, the authorities might wish to control the actual number of immigrants in the country for social reasons.

Consider the first target, the maximization of national income. Measure home real income as: $Y=wL+vS-E$. Here v denotes the wage rate of skilled workers. Then

$$(2)\quad dY/dE = (L+S[dv/dw])(dw/dE)-1.$$

Now if both factor markets clear, we must have $dv/dw=-u$ so that $dY/dE=(L-uS)(dw/dE)-1=-I(dw/dE)-1$. Since $dw/dE<0$, national income must fall. If, on the other hand, the unskilled labor market has a rigid wage w_U, an increase in E will have no effect on w_U, L, v, or S, but will raise e (benefiting unskilled workers) and E (harming skilled workers). National income could conceivably rise in this case because of the distorted labor market: increased interdiction reallocates scarce jobs from illegal immigrants to legal residents and this has a positive effect on income. In this case $dv/dw=0$ and $dw/dE=w_U de/dE$. From (1) and (2), national income will in fact rise if and only if:

$$(3)\qquad \varepsilon_B > [1-g]E/[w-w^*]L,$$

where ε_B denotes the elasticity of the border enforcement schedule (Eg'/g).

This analysis should leave us skeptical about the ability of interdiction to increase national income. The full burden falls on ε_B. But skepticism about the ability of marginal changes in E to have significant effect is widespread in practice. Furthermore, in this model national income can be raised only by redistributing income from skilled to unskilled workers. With this distribution itself a policy target, there is no freedom to influence national income even if it is feasible to do so. Finally, it is the presence of a distorted home labor market that makes it possible to use E to raise Y. One would expect, therefore, that this possibility would disappear if instead w were to adjust to clear the market for unskilled labor.

We still have two potential targets, e (or the internal *ex ante* distribution of income) and I. But note that with u set by the rigid wage, e directly determines I: the two targets are bound together. Unless the authorities just happen to desire a consistent combination of e and I, they will have to trade off one goal for the other, even before considering what tools might be available.

The discussion thus far furnishes two good reasons to hunt for an additional policy tool to supplement interdiction. First, border enforcement is quite probably costly. Public debate in the United States has for years been dominated by the belief that a border enforcement policy consistent with national goals would in fact prove far too costly. The second reason is that the authorities are likely to have independent goals regarding both internal income distribution and the volume of immigration. Thus there is need for an instrument capable of unbundling these two targets.

The rest of this paper concerns two additional instruments. The first consists of sanctions against employers of illegal immigrants. Such sanctions seem very promising in regard to the two reasons just discussed for looking at additional tools. First, as suggested by frequent public suggestions that illegal entry attempts will abate once the employment prospects of immigrants are reduced, sanctions promise to exert an effect independent from that of border enforcement. Thus a combination of such policies could potentially reduce costs. Second, if the employment of illegal immigrants is prohibited, and if this prohibition is backed up by some degree of enforcement, firms will no longer see the two types of workers as

identical. Thus such a policy fosters the hope of unbundling the two goals of internal income distribution and the volume of immigration. The third instrument I shall consider is amnesty: legalization of the status of some illegal immigrants.

There are also practical reasons for considering the two additional policies. Public debate in the United States has for several years focused upon the advisability and consequences of the adoption of such policies. Also they are important in many northern European countries, where only modest efforts are devoted to border enforcement.

A crucial consideration is whether domestic firms are able to distinguish legal workers from illegal immigrants. If they cannot, the policy will disadvantage all unskilled workers relative to skilled workers. For the firm that hires one of the former takes the risk that the new employee is an illegal immigrant and that the authorities will discover this, thereby increasing the expected cost of employing him. But if firms can distinguish illegal immigrants, domestic enforcement will disadvantage these laborers relative to native unskilled workers and legal migrants. Thus the effect upon this latter group is sensitive to the ability of firms to distinguish between the two types of potential employees. Legitimate unskilled workers then have an incentive to adopt measures to distinguish themselves from their illegal rivals, and to support public policies designed to produce such distinctions.

It is not surprising, then, that such policies should be central to public debate on immigration enforcement in host countries. Suppose that some policy attempts to distinguish illegal migrants from legitimate unskilled workers. This presumably involves "doing something" to the latter, since the former would not cooperate with such an effort. Think of the policy as one of issuing identification cards to legal workers.

The unskilled labor force consists, as before, of illegal immigrants I plus legal unskilled workers L who would never be mistaken for illegal entrants, but now there is also a group N of legal unskilled workers who could be so mistaken. Thus the total unskilled labor force is $U = L + N + I$. Let p denote the probability that a member of I succeeds in passing himself off as a member of N.

As before, a potential migrant who attempts entry faces the probability g of failing, and thus earning $(w^* - k)$, and the probability $(1-g)$ of gaining entry. In the latter event, he faces the conditional probability p of successfully passing himself off as a member of N, and thereby earning their wage w_N, and the conditional probability $(1-p)$ of being recognized as a member of I and earning their wage, w_I. He also faces the probability α of legally becoming a member of N. Assuming as before that attempted migration adjusts to equate the expected reward of attempting entry to that of staying behind,

$$(4) \quad w^* = (w^* - k)g + (1-g) \times \{[w_N p + w_I(1-p)](1-\alpha) + \alpha w_N\}.$$

Now consider what firms would be willing to pay to hire the various types of unskilled workers. (From now on I concentrate on the case where all labor markets clear; it is straightforward to incorporate various types of wage rigidities.) Firms would be willing to pay clearly legal workers the value of their marginal products:

$$(5) \quad w_L = f'(u).$$

When deciding whether to employ workers that are clearly illegal, firms will now take into account the chance that sanctions might be levied against them. Suppose that, if caught, a firm must pay a penalty of k^* for each illegal alien employed. I assume that the fraction of employed illegal aliens discovered by the authorities depends upon the resources, D, devoted to enforcement. Let this fraction be given by $h(D)$, where $h' > 0$, $h'' < 0$, $h(0) = 0$ and $h < 1$. If risk-neutral firms treat h as the probability that the employment of an illegal alien will be discovered, they will be willing to pay

$$(6) \quad w_I = f'(u) - h(D)k^*.$$

Finally, when employing members of the apparently legal group N, firms will realize that they may be dealing with disguised illegal aliens. The fraction of apparent members of N who are not actually members is given by $pI/(pI+N)$. Thus w_N must satisfy[4]

$$(7)\quad w_N = f'(u) - [pI/(pI+N)]h(D)k^*.$$

Equilibrium is determined relative to a chosen policy mix. The variables E and D reflect border interdiction and employer sanctions, respectively. What about amnesty? It is not simply the manipulation of I, because I is determined endogenously and we want to allow the possibility that amnesty might produce an offsetting change in immigration. Thus it should instead be modeled as a change in the number of legal workers. But which legal workers: N or L? This is not clear. If the distinction between these latter two groups is basically ethnic, it might seem better to model amnesty as a shift of illegal immigrants into N. But presumably there will be some systematic (and publicly known) criterion to determine which immigrants receive amnesty. This might well identify those individuals to prospective employers. If so, they should be thought of as moving into L. Both possibilities deserve to be considered. In addition, granting amnesty once could inspire in prospective migrants the hope that it might be granted again. That is, α might increase. In sum, amnesty can be modeled as the choice of an increase in N (and/or in L) accompanied by an induced increase in α.

In this model, policy (and perhaps history) determines the values of E, D, N, L, and α. Equations (4), (5), (6), and (7) can then be solved for the three wage rates (w_L, w_N, and w_I) and the volume of illegal immigration I —recall that $u=(L+N+I)/S$. These in turn imply the values of the other variables of interest, such as the wage paid to skilled workers and the rate at which they must be taxed to pay for the chosen policy.[5]

The model can be used to analyze the implications of alternative policies. It is straightforward to differentiate equations (4)–(7) with respect to the policy variables. So I omit technical detail and merely describe some results.

An increased interdiction effort E reduces illegal immigration, depending upon the elasticity ε_B, and also upon the ability of migrants to obtain legal treatment, that is, upon p and α. The lower I increases the marginal product of unskilled labor and so causes a rise in w_L and in w_I, in equal amounts. But the real winners are the members of N (and those who pass themselves off as members): w_N rises not only because of the increase in marginal product, but also because the reduction in I reduces the risk involved in hiring a member of N. Thus the presence of sanctions increases the benefit to this group of interdiction.

An increase in the enforcement D of employer sanctions does indeed reduce w_I by increasing the risk to employers of hiring illegals, and this in turn induces a reduction in I. Members of N benefit from the resulting rise in the marginal product of unskilled labor and also from the fact that the lower I makes it less risky for firms to hire them, but the latter is more than offset by the direct increase in risk caused by the larger effort D. On balance w_N must rise. But the big winners are the L workers who reap the undiminished rewards of higher marginal productivity. (The fact that the advantage to being thought part of N relative to being realized as part of I has increased will cause migrants to attempt to increase p and α if they can. This could cause sanctions to appear to be more effective in limiting immigration than they actually are.) Clearly using E and D together allows greater control over the unskilled-labor wage structure, together with I, than would either individually, and also any

[4]Note that this formulation assumes that the inspecting authorities are able to identify an illegal alien even if firms cannot. Relaxation of this assumption has little effect, provided that there is some chance that an illegal alien unknowingly employed by a firm will be uncovered and sanctions applied.

[5]The penalties k and k^* could also be policy variables, but I do not consider this.

specific target for I can be achieved at a lower total cost $D+E$ with a combined policy than with either individually.

Suppose that amnesty involves increasing L. A lower I then reduces the risk for employers to hiring members of N and so exerts upward pressure on w_N. This induces immigration, so the unskilled labor supply increases and w_L must fall, as does w_I. Thus amnestied workers are at least partly replaced by new illegal immigrants. If α rises, these effects are reinforced, and a sufficient rise in α could cause I to actually rise and w_N might conceivably fall. If amnesty instead involves a rise in N, this provides a larger direct reduction in the risk to firms of hiring apparent members of N. This implies an even greater initial upward pressure on w_N and so stronger implications.

Note that in the absence of employer sanctions amnesty simply reduces I by the same amount as N (and/or L) is increased. With migrants able to enter the labor force on the same basis as legal workers, any induced rise in α is of no economic significance. The presence of employer sanctions is necessary for amnesty to be a substantive policy.

This is my "first-pass" at modeling illegal immigration. It can be elaborated in various straightforward ways. For instance, another input could be added to allow complementarity between some factors, another sector could be included to reflect the fact that immigration is much more important in some parts of the economy than in others, a period structure could be added to make the dynamic part of the problem explicit, and so on.

REFERENCE

Ethier, Wilfred J., "Illegal Immigration: The Host Country Problem," *American Economic Review*, March 1986, *76*, 56–71.

Can Border Industries Be a Substitute for Immigration?

By FRANCISCO L. RIVERA-BATIZ*

The continuous flow of workers from Mexico to the United States has given rise to a wide range of proposals aimed at containing the migratory tide. At the top of the list is usually Mexican border industrialization, the incarnation of the view that immigration should be controlled at the border, through the creation of an economic fence absorbing the potential migrants. Border industrialization received a strong boost in the mid-1960's with the establishment of the Mexican Border Industrialization Program (BIP). The BIP combines trade, fiscal, and other incentives to encourage the formation of assembly plants (maquiladoras) in Mexico, particularly in the border region. Though the BIP is only part of a wider border industrialization process, it has been particularly connected to U.S. immigration policy. Indeed, the BIP was established in the aftermath of the bracero program (that ended in 1965) and its main purpose was to absorb the former braceros into the Mexican labor force so as to prevent their illegal migration to the United States. In this paper, I examine the impact of the BIP on immigration to the United States.

I. Assembly Plants, Wages, and Employment in the Border Region

The significance of illegals in Mexican-U.S. migration suggests that the key determinant of U.S. nonrefugee immigration is the socioeconomic incentive to move and not immigration quotas, which are not binding. How do border industries affect the incentives to migrate to the United States? The conventional wisdom associates the assembly plants with rising employment opportunities and improved standards of living in Mexico, both of which presumably reduce migration to the United States. What are the facts in this respect?

Consider the impact on wage rates. As in most other Latin American countries, wage rates in Mexico are highly controlled by the government, and organized labor is also closely involved in wage setting, with the government often responding to pressures from the Confederation of Mexican Workers (CTM) to raise administered wages. Maquiladora workers, however, are usually unorganized, part of the package that attracts assembly plants to the area. There is a limit, though, to the downward wage flexibility accorded these firms: the Mexican Minimum Wage Commission sets minimum wages according to type of job and geographical region. Though violation of minimum wage laws is hard to detect and enforcement more difficult among domestically owned plants, foreign-owned firms are usually more closely monitored and less likely to violate the labor laws. Since the majority of the maquiladoras in the border region are U.S.-owned or controlled, one can expect them to pay close to, or above, minimum wages.

Surveys of maquiladoras in Mexico show that, under usual circumstances, they will offer wage rates in the range of 5 to 30 percent above the corresponding official minimum wages (see Michael Seligson and Edward Williams, 1981, pp. 46–49, and Joseph Grunwald and Kenneth Flamm, 1985; Monica Gambrill, 1981, partly contests these conclusions). Since minimum wages are higher in the border region relative to the rest of Mexico, these represent above-average wages in Mexican manufacturing. They do not represent, however, a significant reduction of U.S.-Mexico real wage differentials. By moving to the United States, illegal migrants could obtain as much as an 800–1300 percent increase above the Mexican minimum wages, depending on occupa-

*Visiting Associate Professor, Regional Science Department, University of Pennsylvania, Philadelphia, PA 19104. I have benefited from comments by Michael Piore.

tion, region of employment, etc. Furthermore, with minimum wages not adjusted to keep up with inflationary pressures, the differential between Mexican border real minimum wages and wages available in comparable occupations in the United States has not increased to any perceptible amount since the establishment of BIP and has declined in recent years, particularly if the occupations are unskilled, as most maquiladora employment is (real hourly wages paid to production workers in assembly plants have declined by more than 25 percent since 1979).

What about employment creation? One of the standard criticisms of industrialization initiatives in Latin America has been the perception that the industries established are highly capital- and skilled-labor intensive, providing scant gains in terms of unskilled employment creation. This is clearly not true of assembly plants. Grunwald-Flamm report that in 1979, Leopoldo Solis calculated that capital per worker in a sample of assembly plants was six times lower than that in an average of all Mexican manufacturing; more recent data suggests a narrowing of this differential to three, still a remarkable difference. Indeed, the BIP has been extremely successful in terms of increasing employment in the border region. From 1969 to 1985, the employment of the assembly plants in Mexico grew from 15,000 to 240,000. It is projected that by the end of 1986 more than 800 assembly plants employing close to 300,000 workers will be in operation along the border; this makes the assembly industry one of the major employers in the area.

The fact that the BIP has provided the border region with high rates of employment growth at wages above the Mexican average has led to the conclusion that "by encouraging migration from more remote areas of Mexico to the border, these programs may actually have encouraged illegal immigration to the U.S." (Pastora San Juan Cafferty et. al., 1983, p. 192). The border region's population has indeed grown enormously over the last 40 years, at a much higher rate than other regions. And the main factor behind this growth has been interregional migration, from the south to the north. It has been reflected in high rates of urban growth. For example, the rate of growth of population from 1970 to 1980 in Tijuana was 96 percent; in Mexicali: 85 percent; Ciudad Juarez: 67 percent; Reynosa: 75 percent, and Matamoros: 87 percent, while the national average rate of urban growth was 37 percent (see Niles Hansen, 1985). It is clear that a large share of the border's population growth serves as a pool from which illegal migrants to the United States are drawn. Can we, as Cafferty et al. note above, associate a significant part of this growth with the BIP?

Some of the available evidence suggests that the BIP "contributes only marginally to internal migration to northern Mexico and international migration to the U.S." (Seligson and Williams, p. 169). The main explanation behind this conclusion is the fact that the employment practices of the assembly plants screen out as potential employees most of the labor force that migrates interregionally within Mexico and/or that is likely to migrate illegally to the United States. The maquiladoras' employees are overwhelmingly border-area born, many of them local female workers entering the labor market for the first time. Over 70 percent (and at times close to 90 percent) of the maquiladora labor force is female. As is well known, illegal migrants to the United States are predominantly male workers. Thus, the assembly plants do not offer a viable job alternative for them.

Furthermore, the maquiladoras currently offer very few forward and backward linkages to the Mexican economy. They specialize in importing most, if not all, of their raw materials and parts from the United States, exporting the finished product out of Mexico. This means that the main indirect employment creation of the maquiladoras in Mexico occurs through the services that emerge to cater to the demands of the assembly plants and their labor force. Though some job creation has occurred through these means, a large leakage occurs because of spillovers across the border. A large fraction of the income of maquiladora employees, for example, is spent across the border in U.S. goods and services (see Seligson and Williams). One can conclude that, in its present condition, the impact of the BIP appears to be minimal

in terms of either attracting workers to the border, or of reducing incentives for illegal migration to the United States. The growth of the border area population cannot be attached in any significant way to the border assembly industry.

The situation, however, appears to be changing. Not only do the more recent assembly plants established in the border have higher capital-labor ratios, they also use higher male-female labor ratios. This means that the turnover rate for male workers in these firms must be low enough to make them competitive with female workers. Assembly plant managers frequently justify their hiring of females on the basis that, unlike male workers, they are less likely to quit their jobs once they have accumulated enough to pay for the costs of migrating illegally to the United States, that is, paying the smugglers (coyotes) and temporary living expenses. It is possible that the more recent assembly plants are supplying jobs with some advancement opportunities and with nonmenial tasks, making them somewhat differentiated from the kind of employment provided by traditional maquiladoras. They might thus be providing attractive employment opportunities to potential migrants, discouraging those workers with relatively high aversion for illegal migration from moving further north. Whatever the reason for the employment shift, and assuming it continues, would it help in reducing migration to the United States?

II. Employment Creation, Expectations, and Labor Migration

The impact of the maquiladoras on migration to the United States is dependent on whether they are able to raise employment by more than they increase the labor supply in the border area through induced migration from southern regions. If an excess supply of labor is generated at the border, with surplus workers becoming either openly unemployed or underemployed, it is likely that there will be an spillover into illegal migration to the United States.

It is in the nature of the migratory process, particularly in developing countries, that it involves key speculative aspects. Though information about employment opportunities in destination areas is usually available—through networks of friends, relatives, etc.—it is impossible to know in advance exactly how many positions will develop or how many applicants will seek employment. This is particularly so when the distances involved are large, as they are in the case of interregional migration. In addition, since migration will usually be undertaken for an extended period of time (depending on the temporariness of the migration), there is an intertemporal aspect involved: expectations must be formed not only about current but also about future employment prospects, particularly if the worker is participating in unskilled labor markets with relatively low rates of tenure. Expectations formation, then, is central to determining the impact of job creation on induced labor flows and, as a consequence, unemployment and/or underemployment.

Suppose, for simplicity, that production in the border region, over any given time period t, is undertaken through the use of capital, K_t, and labor, N_t, by means of a Cobb-Douglas production function $Y_t = AK_t^a N_t^{1-a}$, where A and a are parameters representing the (fixed) technology used. In determining the demand for labor, firms set marginal productivity equal to the own-price real wage in the border, W_t^d, or in natural logarithms

$$(1) \quad \log W_t^d = \log A(1-a) - a\log N_t + a\log K_t.$$

Labor supplied to the border is assumed to take a log-linear form

$$(2) \quad \log L_t = b_0 \log\left[\left(W_t^d N_t/L_t\right)/W_t^0\right] + \sum_{k=1}^{T} b_k \log\left[\left(W_{t+k}^d/W_{t+k}^0\right) \times E\left[N_{t+k}|I(t)\right]/E\left[L_{t+k}|I(t)\right]\right],$$

with L_t equal to the labor force in the border region, and W_t^0 the wage rate in the rest of Mexico. The first term in the right-hand side

of equation (2) shows the influence of the differential in expected earnings in period t between the border and other regions on the labor supplied to the border. It is assumed that the probability of employment in the border is, à la Harris-Todaro, given by the employed labor force divided by the available labor supply in the region. Though this represents a situation where all jobs turn over and are randomly allocated among available workers in each time period, it greatly simplifies the exposition and could be modified without altering the results. It is also assumed that the probability of employment in the south is one, though the wage differential between border and southern regions is greater than one. The wage differential is assumed to be exogenously determined by the government and expected to change in an exogenous way according to the declared interregional equality goals of the government. The parameter b_0 represents the elasticity of the border labor force at time t with respect to the expected wage in the border relative to that in southern regions at time t.

The second term in equation (2) depicts the influence of the future interregional differentials in expected earnings (i.e., in periods $t+k$, where k goes up to T, the time horizon) on current labor supplied to the border, L_t. These expected earnings differentials are calculated in a similar fashion as that in period t, except that the future wages and probabilities of employment in the border, N_{t+k}/L_{t+k} are based on the expectations of migrants conditional on the information they have available at time t, $I(t)$. The parameters b_k depict the elasticity of labor supply to the border at time t with respect to $t+k$ expected wages.

Equation (2) is a simplified dynamic labor supply function (of the type discussed by James Heckman and Thomas MaCurdy, 1980), and can be obtained from an intertemporal optimizing model of labor migration. Time preference and discounting are embodied into the b parameters, that are all assumed to be positive. The idea behind the latter is that, everything else held constant, higher relative expected earnings in the border, whether at time t or $t+k$, raise the present value of lifetime earnings in that area, inducing south to north migration and augmenting the border's labor force. The b_k's could, however, be negative if there are strong intertemporal substitution elasticities. Higher expected earnings in future time periods $t+k$ relative to the current time period t could induce some migrants to delay their decision to migrate, reducing the border labor force at time t. I assume this effect is small enough so it does not change the positive sign of the b_k's.

Substracting (2) from (1) and simplifying yields

$$(3)\quad \log(N_t/L_t) = [1/(1+b_0)]\,Z_t - [1/(1+b_0)] \times \sum_{k=1}^{T} b_k \log[E(N_{t+k})/E(L_{t+k})],$$

where
$$Z_t = (1-a)\log A(1-a) + \log K_t - 1/a \log W_t^d - \sum_{k=0}^{T} b_k \log\left(W_{t+k}^d/W_{t+k}^0\right).$$

Equation (3) states that the employment rate in the border at time t is related to a group of exogenous variables at time t, denoted by Z_t and to the future expected employment rates at times $t+k$. The latter affect present (time t) employment rates, since they influence the present value of migrant earnings and therefore the current amount of interregional labor migration to the border. How do we model the formation of expectations regarding future employment rates? A first hypothesis is rational expectations—the expectations of migrants have to be consistent with the model of the border labor market described in equations (1)–(3). Using equation (3), for example, the expected employment rate for period $t+1$ is

$$\log E[N_{t+1}]/E[L_{t+1}] = [1/(1+b_1)]\,E[Z_{t+1}] - [1/(1+b_1)] \times \sum_{k=2}^{T} b_k \log(E[N_{t+k}]/E[L_{t+k}]).$$

Substitution into (3) and forward iteration results in

$$(4)\quad \log(N_t/L_t) = [1/(1+b_0)]\,Z_t$$

$$-[1/(1+b_0)]\sum_{k=1}^{T}[b_k/(1+b_k)]\,E[Z_{t+k}]$$

where, for expositional simplicity, equation (4) considers only the direct effects of the expected exogenous variables at time $t+k+j$ on the employment rate at time t and ignore the, relatively minor, chain effects of $E[Z_{t+k+j}]$ on N_t/L_t by affecting $E[N_{t+k}]/E[L_{t+k}]$.

Equation (4) shows the equilibrium employment rate as a function of a vector of current exogenous variables and of a vector of expected future values of those exogenous variables. Most models of labor migration ignore the role of expectations in shaping the migration decision. This is unfortunate as the mechanics of the process of migration is intimately linked to expectations. For instance, equation (4) shows that current employment rates (and therefore unemployment rates) are affected in different ways by creation of employment opportunities in the border region depending on whether the changes are anticipated or unanticipated. An expected increase in the capital stock in the border at time $t+1$, a rise of $E[K_{t+1}]$ in our terminology, reduces the current employment rate, N_t/L_t. The reason is because, through the process of interregional migration, the relative labor force in border areas is positively affected by expectations of future employment in the region. The anticipated capital accumulation does not increase employment immediately, though, generating in the short run only a rise in the unemployment rate. In terms of equation (4), if there is anticipated capital accumulation in period $t+1$, the impact on current unemployment—as a proportion of the labor force—is, everything else constant, equal to

$$dU_t/L_t = d\log L_t$$

$$= [b_1/(1+b_0)(1+b_1)]\,d\log E[K_{t+1}] > 0,$$

where, for benchmark purposes, it is assumed that there is full employment initially ($L_t = N_t$), and where $dN_t = 0$ since the employment disturbance occurs in period $t+1$. The effect is particularly strong if the anticipated employment creation is permanent (so that all $E[K_{t+k}]$ rise) and not transitory (as in the case of only a rise in $E[K_{t+1}]$).

Unanticipated new border programs or investments, such as an unanticipated increase in the capital stock at time t, K_t, do not tend to increase short-run unemployment rates in the border as expected disturbances do. Algebraically, the impact on unemployment is given by $dU_t/L_t = -[1/(1+b_0)]d\log K_t$, which is negative, meaning that unemployment, as a proportion of the initial labor force, declines in response to the unanticipated disturbance. This is diametrically opposite to the case of an anticipated job expansion, in which unemployment increases, and is strengthened if the unanticipated disturbance is temporary and the anticipated one permanent. The explanation is that unanticipated disturbances do not have the speculative impact that anticipated disturbances have. In the case of unanticipated capital accumulation, responses are made on the basis of created jobs not anticipated ones.

This is a key point when expectations are not realized, which is likely to arise in the presence of forecasting errors (under uncertainty), or when expectations formations are based on inaccurate information or extraneous beliefs, deviating systematically from fundamentals (for a discussion of these deviations in financial and foreign exchange markets, and their relation to speculative bubbles, rational or not, see Olivier Blanchard, 1979; Rudiger Dornbusch, 1982). Within the present context, unrealized anticipations of future employment in the border would cause migrants, many of whom have moved there on the basis of those expectations, to face unanticipatedly high probabilities of unemployment and/or underemployment in the area. Some would then be forced to migrate to the United States. This is not a trivial matter as the maquiladoras are highly mobile firms and can switch locations with relative ease. Border employment creation could therefore be highly sensitive to unanticipated events curtailing Mexico's

competitiveness in attracting assembly plants, and/or to business cycles.

My discussion suggests that border industrialization, when expected to be sustained over time, and if a feasible alternative for migrant workers, is likely to generate widespread interregional migration to the border area and could potentially raise the level of unemployment and underemployment there in the short run. Similar behavior is known to occur in relation to urban job creation and rural-urban migration and, as I have described, can be explained by the speculative aspects of the process of migration.

III. Concluding Remarks

In sum, I hope to have thrown strong doubts on the view that the Mexican border assembly industry reduces labor inflows to the United States and, on the contrary, shown that border industries could actually encourage such migration. But even if border industrialization could reduce migration to the United States, there remains the question of whether it would be to the national advantage of the United States to encourage it by stimulating, say, direct foreign investment in the area. In other words, could not such a move be counterproductive, meaning that immigration would provide greater economic benefit (or lower net loss) to the United States? My companion paper (1985) has dealt in detail with this issue.

REFERENCES

Blanchard, Olivier J., "Speculative Bubbles, Crashes and Rational Expectations," *Economics Letters*, 1979, 387–389.

Cafferty, Pastora San Juan et al., *The Dilemma of American Immigration: Beyond the Golden Door*, New Brunswick: Transaction Books, 1983.

Dornbusch, Rudiger, "Equilibrium and Disequilibrium Exchange Rates," *Zeitschrift fur Wirtschafts-und Sozialwissenshaften*, September 1982, *102*, 573–99.

Gambrill, Monica C., "Employment in the Maquiladoras, the Case of Tijuana," (transl.), *Papers and Proceedings* of First Conference on Regional Impacts of U.S.-Mexico Economic Relations, July 1981, *2*, 245–59.

Grunwald, Joseph and Flamm, Kenneth, *The Global Factory: Foreign Assembly in International Trade*, Washington: The Brookings Institution, 1985.

Hansen, Niles, "The Nature and Significance of Border Development," in Lay James Gibson and Alfonso Corona Renteria, eds., *The U.S. and Mexico: Borderland Development and the National Economies*, Boulder: Westview Press, 1985.

Heckman, James J. and MaCurdy, Thomas E., "A Life Cycle Model of Female Labor Supply," *Review of Economic Studies*, January 1980, *47*, 47–74.

Rivera-Batiz, Francisco, "Is Direct Foreign Investment a Substitute for Labor Inflows?: A Review of the Issues and Policy Implications," paper presented at the Western Economic Association Meetings, Anaheim, July, 1985.

Seligson, Mitchell A. and Williams, Edward J., *Maquiladoras and Migration Workers in the Mexico-United States Border Industrialization Program*, Austin: University of Texas Press, 1981.

Government *R&D* Programs for Commercializing Space

By LINDA R. COHEN AND ROGER G. NOLL*

Beginning in the late 1950's, the federal government has supported several programs to develop commercial uses of space. Early on, the National Aeronautics and Space Administration (NASA) supported the development of several generations of communications satellites. In the 1970's, NASA developed the Space Shuttle which, although officially declared operational, still consumed much of NASA's budget in the 1980's.

These projects exemplify a class of programs in which the government seeks to advance technology to serve specific commercial objectives. NASA may regard exploring and using space as a valid national goal, but both of these projects were justified externally on economic grounds—both would reduce costs of existing activities, and make possible new commercial activities.

The two programs met different fates. The communications satellite program, after producing several important commercial advances, was killed in 1973. The Space Shuttle, despite cost overruns and performance underruns, survived political challenges and appears to be permanent. The purpose of this paper is to apply recent developments in the theory of policy decisions by elected officials to illuminate the adoption, implementation, performance, and ultimate fate of these programs.

I. Normative and Positive Theories of Government *R&D*

The economics of technological change provides a rationale for *R&D* programs to develop a specific commercial technology. First, the standard appropriability arguments—that nonappropriability may cause underinvestment and duplication in *R&D*, and that the cost of greater appropriability may be monopoly—apply in varying degrees to different industries and technologies. Second, the minimum efficient scale of *R&D* may be so large that efficient private *R&D* requires imperfect competition in the product market. Third, government regulation, taxation, procurement, and other programs may create barriers to innovation. Fourth, if government is a major user of a product, it may vertically integrate into *R&D* to direct technological change towards its own specifications or to appropriate the gains from technical progress.

Our concern here is not with the validity of these arguments, although for space they are not frivolous. Telecommunications is regulated and, internationally, is quasi nationalized, and both policies probably distort the rate and direction of innovation. Government uses satellites and the Shuttle so that the procurement rationale may apply. More generally, *ex ante* evaluations of both programs predicted positive net economic benefits even though private development was not being pursued. We uncritically accept this economic case to address another issue: Do these rationales, if present, comport with the incentives facing political officials who adopt and implement such programs? Does the political system induce political leaders to translate valid efficiency rationales for government-supported commercial *R&D* into an effective policy response?

†*Discussants*: Joel Scheraga, Rutgers University; Marcellus Snow, University of Hawaii.

*University of Washington, Seattle, WA 98195, and Stanford University, Stanford, CA 94305, respectively. We gratefully acknowledge the research support of The Brookings Institution, the Caltech Energy Policy Studies Program, the Center for Advanced Study in the Behavioral Sciences, the Guggenheim Foundation, the Hoover Institution, and the National Science Foundation, none of which bears any responsibility for the contents of this paper.

Since the early 1970's, substantial progress has been made in characterizing the incentive structure of the American political system, especially with respect to legislators. This work assumes that political leaders are self-interested utility maximizers. Their objectives may be a complex amalgam of income, power, and programmatic outputs, but a necessary condition for obtaining the benefits of office, which accrue over several terms, is to survive successive reelection campaigns. Hence, the probability of reelection looms important in maximizing the present value of a legislative career. Reelection prospects for legislators partly depend on the visible, accountable benefits they deliver to constituents. For the president and bureaucrats, reelection has less direct importance; however, serving the reelection needs of legislative supporters contributes to obtaining the programs and budgets that executive officials desire.

The continuing presence of impending reelection campaigns has several logical consequences regarding the behavior of government officials in adopting and implementing *R&D* commercialization projects. Here we summarize the key points that apply to the space program (see our forthcoming study for more details).

Impatience. Legislators face tradeoffs among programs with different intertemporal distributions of costs and benefits. Some, like *R&D*, require expenditures for several years before benefits accrue. Electoral incentives bias the mix of government programs away from long-term projects by inducing a rate of time preference that exceeds the private rate of government officials. Public officials cannot make credible long-term commitments about programs (see Morris Fiorina, 1981). With nonzero probability, a long-term project will be canceled before completion, regardless of its economic success, owing to changes in the composition of the legislature, the identity of the president, or the general state of national politics. Hence, in evaluating a proposed program, voters and legislators multiply future expected benefits by the probability that the program will be completed. This is tantamount to applying an additional risk discount beyond the economic and technical risks of the project. Elected officials further discount future benefits by their probability of reelection, for they cannot collect the political benefits of a program unless they retain office.

Impatience affects implementation as well as adoption. Projects normally begin with low-cost research activities to explore technical alternatives before a design commitment is made for the more expensive stages of prototypes and demonstrations. Impatience leads to decisions to foreshorten research, committing to a design earlier than is optimal. This makes long-term projects subject to cost overruns and performance underruns in relation to *ex ante* engineering and economic estimates that assume optimal project management and use standard discount rates.

Risk Aversion. In American legislative elections, incumbents are overwhelmingly likely to be reelected. Because reelection probabilities are bounded above by one, the relationship between reelection changes and programmatic actions eventually must be concave. Even if a legislator is personally risk neutral, the expected benefits function eventually must exhibit risk averseness with respect to program performance. Congressional scholars observe that legislators obtain high probabilities of reelection by sticking to relatively uncontroversial activities: constituency service and advocacy of only widely popular programs among constituents (Richard Fenno, 1978; and Fiorina-Noll, 1979). The major exceptions involve highly salient national issues or "crises" that force candidates to take positions or suffer electoral setback.

Because *R&D* is inherently risky, legislative risk aversion should induce a bias against it. Among *R&D* projects, the most favored should be closely associated with currently salient issues, or for which government actions can mitigate the risks of political failure. If the government is a major user of a technology, public officials can influence its apparent success by appropriating funds to force its commercial adoption. Legislators can also mitigate risks by linking a project to noneconomic objectives towards which progress is difficult to measure compared to commercial objectives.

Costs as Benefits. After a program is enacted, legislators influence where the funds will be spent to undertake it. Contracts and jobs are not controversial: even opponents of a program want their representative to help local firms acquire program expenditures. This enables effective legislative facilitators to convert some program costs to political benefits (Barry Weingast et al., 1981). The structure of Congress enhances the importance of facilitation activities compared to programmatic decisions. The latter are fragmented among the president and a large number of legislators, but, for each voter, facilitation responsibilities are shared by only one Representative and two Senators. Normally voters can rationally conclude that their representative was not decisive in a program's adoption, but was important in channeling expenditures towards the home district.

The *R&D* programs deliver few distributive benefits when enacted. Local expenditures must exceed a threshhold of importance to be politically significant, and early research is usually inexpensive and spread among several projects. Future, more important expenditures may not be politically important *ex ante* because the site of the prototypes and demonstrations is unknown while alternative designs are under consideration, and because voters and legislators will heavily discount future plans due to impatience. A project's distributive aspects become politically more important as expenditures rise and distributive uncertainties are resolved. The growing importance of distributive benefits can make a project more attractive politically as it progresses, even if its performance is deteriorating. Thus, foreshortening research to arrive quickly to the more expensive phase may be politically beneficial even though it is economically and technically suboptimal.

Distributive effects can work against a program that is otherwise successful. The public benefit of a project is determined by its effects on industry performance, but specific firms experience private effects that depend in part on whether they participated in the program. Specifically, firms that receive contracts may acquire technological advantages over competitors without contracts. Legislators who represent losers will regard this as a distributive disadvantage. To avoid this source of conflict, government must spread contracts among most competitors or centralize the project in a government lab or an industry joint venture. Fragmentation is likely to be more attractive in concentrated industries (such as aerospace), but efficiency considerations are likely to make centralization more attractive in atomistic industries (such as agriculture). When both are too inefficient to be attractive, an otherwise promising program can become politically unacceptable because its effective execution requires concentrating the industry by advantaging a minority of its firms.

II. Applications to the Space Program

The preceding analysis supports several conclusions about *R&D* programs that are consistent with the history of space commercialization projects. The principal predictions are as follows:

1) The public sector is reluctant to undertake warranted commercial *R&D* unless it is attached to a salient national political issue or is directed at the production of government goods.

2) Commercial *R&D* programs are prone to cost overruns and performance underruns because of early termination of research, especially in programs of long duration, and the political attraction of large expenditures in later stages of a program.

3) The net benefits of a program are likely to play a more important role in decisions to adopt a program and to move on to development; however, as expenditures rise, distributive issues become more important.

4) The chances for political survival of a program decline if the target industry becomes more competitive and if the project cannot reasonably efficiently be fragmented or centralized.

Two space programs cannot establish the validity of these propositions. Here we show that they appear to adhere to these principles. For more elaboration, see our forthcoming book.

Communications Satellites. After the Sputnik launch in 1957, the American space

program became a salient issue because of its connections to national prestige, defense, and political competition with the Soviet Union. The federal government responded by allocating more resources to space. One new initiative was the Applications Technologies Satellites (ATS) program to develop and to demonstrate new commercial uses for satellites. The program consisted of a series of relatively short projects, with launches every few years, and was widely acclaimed as successful. Among its technical achievements were directional antennas, advanced satellite stabilization techniques, broadcasts to low-cost ground receivers, and the use of high-frequency portions of the electromagnetic spectrum.

In 1973, ATS was canceled as part of a general reduction in space activities. The official reason for the cancellation was that the satellite industry was now commercially viable and could advance technology without public support. Subsequent assessments of private satellite *R&D* contradict this judgment and conclude that the ATS cancellation stimulated foreign programs and undermined U.S. technological dominance. Here the relevant issue is the technical basis for termination in 1973. In fact, nothing had happened technically or economically to alter the program's rationale. The scale and riskiness of research remained large relative to the size of firms in the product industry, and the user industry—telecommunications and broadcasting—remained regulated in ways that inhibited new uses of satellites. Whereas one can question whether the government should ever have embarked on the program, the case for ATS had not diminished.

What had changed in the early 1970's were political aspects of the program—space lost glamour as a salient national issue because the Soviet Union had withdrawn from the competition and because the technical needs of defense had diverged from commercial requirements. In addition, the product market had become competitive, and this created problems for the government contracting process. Only Hughes bid on the first five ATS satellites, but several other companies sought the contract for the last project. A contract dispute ensued, sparking investigations and delaying construction by over a year. Because the losers outnumbered the winners, and because no reasonable alternative could be found to picking one winner from a group of equally able competitors, political actors perceived inequity in the program, and canceled it.

Space Shuttle. For NASA, the attraction of the Shuttle was that it continued a manned program oriented towards an eventual space station and manned mission to Mars. In the executive and congressional budgetary process, these objectives were not adopted and the case for the Shuttle was economic. Buttressed by detailed studies from Lockheed and Mathematica, NASA predicted that the Shuttle would more than offset its costs by reductions in the expense of launches and by making feasible the retrieval of satellites for repair and relaunch, thereby reducing payload expenditures. Approximately half of the benefits were attributable to recovery capability. Most of these depended on the ability to retrieve satellites from geosynchronous orbit, which in turn required the development of the Space Tug. The Tug was to be carried aloft by the Shuttle and flown to geosynchronous orbit where astronauts could repair or retrieve nonfunctioning satellites.

In 1978, the Shuttle ran into technical difficulties. The Tug proved technically infeasible within the time horizon of the program, and then cost overruns began to mount for both construction and operations. The cancellation of the Tug immediately eliminated the expected net benefits of the program, and further problems drove net benefits further negative, so that in about 1978 or 1979 the program should have been killed or redirected (see Jeffrey Banks, 1985).

One plausible redirection was to reverse the commitment to rely on the Shuttle as the dominant launch vehicle for U.S. spacecraft, to stretch out research and development to improve performance and to complete the Space Tug, and to cancel all but one of the orbiters and one launch facility so as to make the Shuttle exclusively a research vehicle. Working for continuation as is were the distributive aspects of the program. Major expenditures were flowing through the con-

centrated space industry, and retrenchment would mean visible losses to contractors. Moreover, the program was nearing its first significant payoff—the first launch in April 1981. Although excluded from the benefit-cost analysis, one public benefit of the program is the consumption value of the launch and astronauts cavorting in space. By the time the crisis peaked in mid-1979, this was but one election away. Thus, the Shuttle survived budgetary attacks after it was understood to be a commercial failure.

III. Conclusions

These two space programs are interesting because of their contrasts. Both are tenuously connected to salient issues: national prestige; defense; international political objectives. One, satellites, was a commercial success but was canceled in midstream. The other, the Shuttle, is an economic disaster but continues to dominate the nation's civilian space effort. The satellite program produced quick payoffs and so escaped the consequences of impatience. But as the industry became competitive, the program sank amidst intraindustry controversy. The Space Shuttle was adopted in the initial euphoria over the Apollo Project, despite its long duration and hence delayed political benefits. By the time its failure became apparent, its distributional benefits were in full force and it could not be slowed, much less killed.

Both programs initially were sold on the basis of their expected net economic benefits, but their ultimate fates were not determined by their performance. Once the financial stakes to contractors and their competitors became large, distributive issues determined whether they would survive.

Obviously, not all successes are terminated, and not all failures are carried to completion and adoption. Our analysis does not even demonstrate that government is less efficient than the private sector in deciding whether to invest in *R&D*, where to place its bets, and how to carry out an *R&D* program. The point is, we believe, more subtle and more interesting than a blanket characterization of the comparative merits of public and private efforts. It is that the political environment of a program is very important in determining whether a project is adopted and successfully managed, and that the process is comprehensible and even predictable. Initially we asked whether political incentives match up with market failure rationales for commercial *R&D* projects. For long-term, risky projects, the answer is no: these characteristics make the government relatively unwilling to undertake projects. The exceptions are programs that are related to salient national issues, such as space in the 1960's or energy in the 1970's, or that produce goods procured by government. Moreover, if a long-term, risky project is undertaken, and then turns out to be less promising than expected, distributive forces work to preserve it as long as it is not seen as inequitable by the contracting industry.

REFERENCES

Banks, Jeffrey, "Political Influence in Government *R&D* Programs: The Case of the Space Shuttle," presented at Annual Meetings of the American Political Science Association, August 1985 (forthcoming in Cohen and Noll).

Cohen, Linda R., and Noll, Roger G., *The Technology Pork Barrel*, forthcoming.

Fenno, Richard, *Home Style*, Boston: Little Brown, 1978.

Fiorina, Morris P., *Retrospective Voting in American National Elections*, New Haven: Yale University Press, 1981.

______ **and Noll, Roger G.** "Majority Rule Models and Legislative Elections," *Journal of Politics*, November 1979, *41*, 1081–104.

Weingast, Barry, Shepsle, Kenneth and Johnson, Christopher, "The Political Economy of Benefits and Costs: A Neoclassical Approach to the Politics of Distribution," *Journal of Political Economy*, August 1981, *89*, 642–64.

Incentive Compatible Space Station Pricing

By JOHN O. LEDYARD*

Space Station, planned to be operational in the mid-1990's, provides an example of the opportunities and difficulties associated with the development of space. This project, currently projected to cost some $8 billion just to reach operational capability, is being designed to achieve several goals, among which are the encouragement of the commercialization of space, the promotion of international relations through the inclusion of international partners, and the continued promotion of space science and technology. Space Station is to be a continuously manned platform in low earth orbit, providing a variety of resources (for example, power, a low gravity environment, manpower, and laboratory space) that can be used by NASA and others in conjunction with their payloads to further these goals. The station is a highly complex, multidimensional *R&D* project for which good management will be necessary if it is to be successful.

In this paper I discuss several possible contracts between NASA and others and their implications for the operation, pricing, and evolution of Space Station as a major space project. Unalterably interconnected with these pricing policies are the allocation of resources produced on Space Station and the extent of the benefits to be received by the users. Although it is sometimes difficult for an engineer to accept, prices will affect behavior and use patterns, which in turn will affect the ultimate gains from any project designed to operate over a long period of time. Modern economic analysis provides us with a way to analyze these implications and evaluate their impact on the desired goals.

The economic theory that provides the most insights is called mechanism theory or the theory of implementation. A brief survey on topics relevant for this paper can be found in Roy Radner (1986). For now, it is important only to understand the general framework of this theory of organizational design. There are two key hypotheses: 1) the information needed to achieve organizational goals is initially dispersed and difficult to uncover through direct monitoring; and 2) individuals will reveal that information and respond to instructions and requests only if it is in their interest to do so. The designer of the institutional rules, that specify who tells what to whom and who carries out what actions, can do nothing about the initial distribution of information or the motives of the various actors in the organization. The designer can only optimize the organizational goals subject to the informational and incentive constraints. But, within these constraints, there may be a wide range of options, some of which are more desirable than others.

To apply the insights of this theory I first describe the Space Station environment (those features of the project which are not really under the control of the designer), then briefly discuss some of the goals which have been proposed, and finally discuss several options for pricing. I conclude with an open question for research.

I. The Space Station Environment

In this section I briefly describe some of the aspects of Space Station that are important from the point of view of the economist as a designer of organizations and which effectively lie outside our control. (Those interested in a more detailed formal analysis should consult my 1984 paper, or Jeffrey Banks et al., 1985.) The key observation is that Space Station is a multiple-product, highly uncertain, public enterprise, among whose clientele are a wide spectrum of users from the public and private sectors.

*California Institute of Technology, Pasadena, CA 91125. Some support from NASA was provided for this research. They bear no responsibility for this paper.

A. *Technology and Costs*

In simple terms, Space Station is a multiproduct public utility. A major initial capital-intensive investment produces an entity which provides a stream of resources over time, requiring relatively low and, probably, reasonably constant per unit operating costs. However, the differences between Space Station and a standard utility are important. On Space Station the technology is not well understood. Uncertainties exist because these technologies have never before been operated in space at this scale. Secondly, the resources that will be produced are also required as inputs: power is not only to be supplied to users of Space Station, it is also a required input to the life-support and command systems. Since there are significant uncertainties about how many of these resources will be needed for internal use, there are derived, magnified uncertainties about the net amounts available to users. These housekeeping needs should be known once Space Station is fully operational but, for all contractual agreements made prior to the mid-1990's, this is a major uncertainty. Similarly, because this technology is new, there is a lot of uncertainty about the costs of construction and operation. This is, therefore, a large complex project for which comparable endeavors are difficult to find. Standard public enterprise or regulated utility models in which there is a fair degree of certainty about the technologies, costs, and demands are simply inappropriate as models of Space Station.

B. *Demand and Benefits*

There will be an incredible variety of users of Space Station, but for the purposes of this paper one can think of five main categories: commercial users, NASA science and technology missions, other U.S. government users (mostly Department of Defense), international partners (Canada, European Space Agency, and Japan), and all others. Although each of these user classes presents different problems and constraints with respect to pricing policies, they have one thing in common. Benefits and demand are highly uncertain to them and to the designers and operators of Space Station. As economists, we have absolutely no way to use modern econometric analysis to estimate demands (and therefore, perhaps, to estimate consumers' surplus) as might be done in the design of pricing policies for public electric utilities. We also cannot simply ask potential users of Space Station how much of each resource they wish to consume and then plan around the aggregate response. Even if they were certain of their benefits, they have little incentive to reveal all their information. If charges do not depend on their responses, then they have an incentive to overstate their needs; if charges depend on their responses, then they have an incentive to claim only marginal benefits from use. There are no independent market data that NASA can use to check the validity of the data. (Some data are available from STS, the shuttle missions, which with the exception of satellite launches have tended to be short-term research projects flown for virtually no charge. There appears to be very little relationship between these and the long-term projects envisioned for Space Station.) Any pricing rule or other organizational choice must, therefore, not assume that accurate demand or benefit information is available. This immediately rules out a number of policies normally touted in the literature and in the halls of Congress.

II. Pricing Goals

In order to provide a reasonable analysis of alternative pricing policies for Space Station, we need to first consider the goals. What is one trying to accomplish with the pricing policy? As any economist should expect, two desired outcomes are 1) the recovery of some or all of the costs of design, development, and operation, and 2) the Pareto-efficient utilization of Space Station once it is operational. Pareto efficiency implies, for example, that given any particular vector of desired outcomes, the aggregate lifecycle costs of the station and all of its payloads should be minimized subject to achieving those outcomes. This is broader than the goal of minimizing only station costs, but is appropriate from an economy-

wide perspective. Minimization of station costs in a way that imposed significant burdens on the users' costs of building and operating their own payloads would not only be Pareto inefficient, but also may be politically risky for NASA.

Three other goals that may be at least as important as the first two are the promotion of 3) the commercialization of space, 4) science and technology, and 5) international relations. These are important since they relate to the three major user groups: private industry, NASA science and technology missions, and the potential international partners. It is well known that for projects with large set-up and common costs and small operating costs, it is generally not possible to satisfy all five goals simultaneously. For example, a naive approach to achieve goals 3–5 would be to provide station resources free to those users. This obviously conflicts directly with the goal of cost recovery. Less obviously, there is also a conflict with the goal of efficient utilization; too few users will use too many resources. Thus, although a small number of potential users would benefit from a "free access" policy, a larger benefit can be obtained from a more efficient pricing policy. I am prepared to argue that the latter three goals are all advanced if Space Station is utilized and operated in as efficient a manner as possible, and that they are hindered if the pricing policy is inefficient; there is no conflict with the second goal. Efficiency means "more bang for the buck," more resources per dollar input, which means more payloads of all types are able to be accommodated.

The real question, then, is the common one: can the conflicting goals of cost recovery and efficiency be dealt with in a sensible way? For traditional projects with large set-up costs, low marginal costs, and a fair degree of certainty, economists have suggested Ramsey pricing to maximize benefits subject to covering costs. This policy requires either direct knowledge of the demand functions (to estimate consumers' surplus directly), or a tatonnement process with little misrepresentation. Neither option is available for Space Station: demand and benefit uncertainties rule out the former, while cost uncertainties (combined with an application of the Revelation Principle) seem to rule out the latter.

III. Cost Recovery through Posted "Average Cost" Pricing

One simple proposal, intended to accomplish the goal of cost recovery with little damage to efficiency, would be to charge a price for each payload equal to the cost of its STS (shuttle) flight plus a percent to cover the rest of the costs of accommodating it on Space Station. (For now I put aside the problems involved in determining which costs are to be recovered: design and development, construction, launch, and/or operating costs.) This is a reasonable policy only under two assumptions: 1) the designers, builders, operators, and users of Space Station are a team (in the sense of Jacob Marshak and Radner, 1972); they are in agreement about the goals for Space Station and, although perhaps asymmetrically informed, are willing to provide any information they have when requested; and 2) no user will alter his decisions as a result of the prices he is charged (prices only allocate costs, not resources). Both assumptions are false. One need only consider the current pancake-shaped engine for satellites developed by Hughes Aircraft in response to the pricing policy for the STS shuttle to realize that even engineers respond to price incentives. This is especially important for Space Station since that project truly broadens the user classes to include other than U.S. government funded missions.

It is easy to predict what would happen if "shuttle plus a percent" were instituted. Payloads would be designed to conserve on weight and length, but not on either power or manpower, two of the station resources which are projected to be in a constant state of excess demand. Since time on station is not included in the billing calculation, long-term missions requiring a lot of tending by mission specialists will be preferred by the users to brief missions even if the latter use few station resources. Designers are trying to build the best station possible at the lowest possible initial cost. Although they recognize the need to minimize the present discounted

costs of construction and operation, the congressional budget constraints they have been given do not encourage intertemporal tradeoffs. The designers have incentives to minimize construction costs and to hope that operating costs will not be too bad. Robotics are ignored; manpower will do. Methods to conserve consumables will be downplayed; shuttle trips can be expanded. This will not be intentional, simply the rational response to the constraints and pressures placed on the builders. High operating costs mean more expensive, or perhaps fewer, resources once the station is operable. This form of pricing policy leads inevitably to inefficiency, and does not implement most of the goals of Space Station.

IV. Efficiency through Posted "Marginal Cost" Pricing

A natural way to rescue some efficiency is to price each critical resource. This can be done by posting long-run marginal costs, two-part prices, short-run marginal costs, or Ramsey prices, and agreeing to supply demand up to the amount available. Each specific policy yields different implications for the level of benefits obtained and the revenue received, but the posting of prices that approximate expected marginal costs will definitely lead cost-conscious users to design payloads to conserve on those resources whose supply is relatively difficult or costly to expand. But some difficulties exist. The extreme uncertainty concerning costs means that either users will find it difficult to impossible to predict what the marginal cost prices will be, or NASA will have to post an expected value and absorb all the risk themselves. The latter seems to be a no-win situation for NASA. If the posted prices are lower than marginal costs, then Congress or NASA must absorb the loss (causing potential difficulties for future NASA funding), and if the posted prices are higher than marginal costs, then users will be outraged (also causing potential political difficulties). The former is also bad. Uncertainty over prices will cause risk-averse users to postpone development of payloads that should go now. Also, with sufficiently variable supply, even if payload builders feel they can correctly anticipate prices and proceed with their own expensive *R&D* programs, the possibility exists that they may be rationed out in the early years (until appropriate expansion can occur). For small, short-duration payloads this may not be much of a problem; for large, long-duration payloads they should be intolerable and should lead to fewer of this type of mission than is efficient. The standard solution for this problem is some type of insurance arrangement, provided either by NASA or private markets, like that initially available for shuttle launches of satellites.

V. Efficiency through Contingent Contracts

Posted prices, even if they equal long-run marginal cost, may lead to inefficient utilization of the Space Station because of the large uncertainties involved in the new technology and because of the risk attitudes both of potential users, and of Congress and its electorate. When there are large uncertainties about costs or supplies (large as a percentage of the project), not controlled by either supplier or user, and both parties are risk averse, then contingent contracts can make both better off and improve efficiency. In the move which usually works, it is agreed to deliver more of the uncertain output (for less per unit) if supply is large and costs are low, and to deliver less output (for more per unit) if supply is scarce and costs are high. Agreements to supply an amount conditional on total resource availability at the realized long-run marginal cost per unit would be one type of contract like this. There are obviously many others.

The difficulties in writing such contracts arise in identifying appropriate "mutually verifiable (without costly monitoring) and exogenous (independent of each others actions) events" on which to make the contracts contingent. On Space Station the true realizations of both supply and costs are partially under the control of NASA, its engineers, and the builders of the station. For example, if NASA controls all the nonlaboratory aspects of the Station, it has an incentive to include as much of its own power consumption under the housekeeping rubric as it can.

(Power used to support manpower in medical experiments may be accounted as used for life-support systems for the crew.) Cost manipulation like this would be even harder to detect directly. International partners and major commercial users, among others, are suitably skeptical about NASA's ability to keep operating costs (much less construction costs) under control. They need only look at the early predictions of shuttle operating costs and compare them to the actual realizations to support their suspicions.

On the station, almost all easily observable events one could possibly use for contingent contracts are the compound result of the actions of one actor and some exogenous event. These are difficult to separate to the satisfaction of all parties without extensive monitoring and auditing. It is therefore unlikely that these types of contingent contracts will lead to the efficient operation of the Space Station. The standard solution to this moral hazard problem is the use of principal-agent contracts.

VI. Principal-Agent Contracts

If moral hazards exist, but it is possible to organize the project so that the uncertainty is either exogenous or under the control of a single party (the agent), and if that agent is less risk averse than the others (the principal(s)), then the efficient contract consists of giving the agent full control of the project, giving that agent rights to all the uncertain benefits and responsibility for the uncertain costs, and paying the other(s) a fixed fee (possibly negative, possibly in kind) which is not contingent on any of this. As an example, NASA could agree to deliver a fixed vector of resources to each user in return for a fixed payment (to be negotiated). If NASA is risk neutral, this is the most efficient contract; given any other contract, there are terms of exchange for this one such that both NASA and the user are better off.

For large contracts (that involve, say, one-third of the projected power to be supplied), NASA does not appear to be risk neutral. Then the efficient contract involves the agent laying off some of the risk on the principal in return for better terms of trade or a reduced fixed fee. In profit-making operations this can be done through a profit-sharing arrangement (for example, the principal gets 20 percent of the net profits in return for an investment of $1,000), but on Space Station there is no market for much of the output, and therefore any such arrangement must be in terms of the resources to be supplied. A user who projects a large resource demand for a long duration might be willing to take 20 percent of the realized net resources of the station in return for a payment of some fixed amount initially. This type of user would probably be either a private venture capitalist, buying for resale, or a large, non-U.S. government user, such as one of the international partners. Of course, resale would be allowed to improve efficiency. This type of contract readmits the moral hazard problem, but the user accepts this in return for some concessions on price.

There are three other difficulties with principal-agent contracts. NASA cannot keep its savings because of the form of congressional budgeting; this blunts its incentives as agent. Over time, evolution of the station must be managed to contend with new users and new information; an agency contract does not necessarily provide the correct incentives for situations with extensive learning-by-doing. Finally, many aspects of Space Station use rely on inputs from both the user and the Space Station managers. If it is not possible to organize in a way that isolates the effects of each parties' actions from the other, then some form of partnership arrangement is needed.

VII. Partnerships

When the inputs of both parties are needed to obtain an output, and when the inputs cannot be easily monitored or separated from external exogenous events, then principal-agent contracts are inappropriate. In the operation of a lab module it may be impossible to identify the separate contribution of either party to the success or failure of an experiment. Was it the failure of the command crew, under the control of NASA, to maintain "zero g" or did the payload specialist, under the control of the user, unintentionally

"bump" the payload too hard? If one cannot separate these effects, then it is possible that some agreement in which all costs and benefits (resources) are shared (a closely held partnership) may leave both parties better off than under an agency contract.

The major difficulty with partnerships is the "free-rider problem." When separate effects cannot be identified easily, both parties have an incentive to shirk. This is especially true in *R&D* projects like the Space Station where benefits are not marketable. Only if this is a very long-term, repeated relationship so that tit-for-tat arrangements can be implemented do partnerships of this form appear to be viable. It is also helpful if there are some (possibly imperfect) publicly available measures of individual performance. I have come full circle to the team problem of Section I with the added twist of shares in the resources and costs. Little is known about efficient contracts and incentive compatible pricing in this situation. More research is needed.

VIII. Conclusions

Space Station is a complex, multiproduct public enterprise with large uncertainties about the implications of its technology. It is also the potential prototype of many more. Pricing policy will affect the efficiency of use of Space Station and its evolution. But pricing policy is constrained by the location of information and by the incentives of all participants. For small users, such as non-NASA U.S. users and nonpartner foreign governments, efficiency seems best served by either a contract contingent on the realization of long-run marginal cost and net supply, or, if moral hazard is a fear, a fixed-fee contract for the delivery of a fixed amount of resources. For large users, such as the potential international partners, this arrangement would not be appropriate for NASA. In that case it is best to organize so that each individual is responsible for their own actions through a principal-agent contract. (Each user and NASA could be both an agent on one contract and principal on another.) If that organization is not possible, then some form of partnership is probably efficient, the precise nature of which is not yet known. There are organizational solutions to the management problems that have tended to trouble large uncertain *R&D* projects in the past. It is time to be as imaginative and daring with these management solutions as NASA has been with its engineering solutions.

REFERENCES

Banks, Jeffrey S., Ledyard, John O. and Porter, David, "Pricing, Evolution, Design Planning of Space Station under Uncertainty," JPL Economic Research Series No. 22, Pasadena, August 1985.

Ledyard, John O., "Space Station Pricing Options," JPL Economic Research Series No. 20, Pasadena, October 1984.

Marshak, J. and Radner, R., *Economic Theory of Teams*, New Haven: Yale University Press, 1972.

Radner, R., "Decentralization and Incentives," in T. Groves et al. eds., *Information, Incentives, and Economic Mechanisms: Essays in Honor of Leonid Hurwicz*, Minneapolis: University of Minnesota Press, 1986.

Out of Space? Regulation and Technical Change in Communications Satellites

By MOLLY K. MACAULEY*

Regulatory practices by the Federal Communications Commission (FCC) have the effect of rationing the use of a particular resource required for communications satellite technology, the electromagnetic spectrum. Spectrum, or the "airwaves," is the medium over which communications signals such as TV, telephone, and radar travel. Federal government allocation of spectrum among competing services has long been implemented to mitigate the interference that can arise between nearby signals—hence, for instance, the assignment of radio stations to unique regions along the AM and FM dials. That government regulation can and probably does fail to allocate spectrum efficiently, for all the usual economic reasons, has been attested to, criticized, and in turn the subject of proposed reformation in an economics literature both historic (radio spectrum regulation inspired Coase's "theorem") and growing (including work which dates from Harvey Levin, 1971, and references cited therein, to, most recently, Stanley Besen et al., 1984).

Left unaddressed, however, have been the implications of inefficient spectrum regulation for the pace and direction of technical change. Specifically, the problems of static resource misallocation may be compounded by inefficiency in induced innovation (V. Kerry Smith, 1974, 1975; Koji Okuguchi, 1975; Wesley Magat, 1976). If FCC allocations incorrectly signal the true economic scarcity of spectrum, innovation to augment spectrum and other inputs on the basis of relative scarcity may be misdirected, and the overall rate of *R&D* spending may be distorted accordingly.

The effect of government regulation on innovation in communications satellite technology merits particular attention for several reasons. First, a recent FCC ruling will increase the cost of future satellites by requiring them to operate at FCC-mandated minimum levels of *intensity* of spectrum use, on top of rationed *quantities* of spectrum (see *Federal Register*, 1983, para. 69). Second, unlike other uses of spectrum, there is a large public sector component to satellite *R&D* spending that is also likely to be affected by FCC regulation. Undertaken by NASA, current research expenditures on advanced communications satellite technology have been justified in large part by a perceived need to develop methods that use spectrum more intensively (see NASA, 1984, and U.S. Congress, House, 1984). Third, and again distinguishing satellites from other users of spectrum, the inherently global nature of satellite technology renders satellite spectrum allocations a contentious international issue. In particular, developing countries not currently using satellite technology have expressed serious concern about future spectrum availability. In response, technical change economizing on spectrum is frequently endorsed by regulators, and moral suasion is accordingly brought to bear on industry, as an appropriate solution (see FCC, 1985, and U.S. Congress, 1982).[1]

This paper proceeds as follows. Section I tailors a model of induced innovation de-

*Fellow, Resources for the Future, 1616 P Street, NW, Washington, D.C. 20036. Funding from a Resources for the Future "Investment in People and Ideas" award is gratefully acknowledged, as are comments from Jim Stokes, Mike Toman, Paul Portney, and research seminars at RFF.

[1]A related resource required for communications satellite technology is access to the geostationary orbit. Access by domestic industry is also governed by FCC regulation, and public, private, and international *R&D* issues analogous to the issues described here arise in connection with the orbit (see my paper with Paul Portney, 1984). Note, then, that to the extent the orbit is misallocated, recommendations above may be only second best.

signed by Smith (1974) and Okuguchi (1975) to include an input constraint and, as suggested by Morton Kamien and Nancy Schwartz (1969), rates of *R&D* spending. Section II evaluates the model using estimates of the shadow value of spectrum and parameters of an unconstrained production function fitted to engineering data which relate spectrum and other inputs. Results suggest that inefficient spectrum allocations distort both input choices and technical change in directions which vitiate the promulgated emphasis on spectrum *R&D*. Moreover, the bias imparted to factor ratios by misdirected technical change can be quite large. The concluding section adds a refrain to the chorus of arguments mentioned above for improving the efficiency of spectrum use.

I. The Model

In the lengthy literature modeling induced technical change, two factors in particular have been identified as influencing the direction and magnitude of innovation. The first influence is relative input prices; all else equal, firms have an incentive to find new ways to economize on more expensive inputs. Second, the output of the *R&D* process is typically represented by an innovation possibilities frontier that is akin to a concave production possibilities frontier, but with measures of factor augmentation on the axes. The effect on production of selecting from this menu is to augment inputs according to the opportunities traced by the frontier. The decision of where to locate along the frontier depends on relative factor shares in the total costs of production; it is important to note that because factor shares also depend on the isoquant relating augmented factors, choices of input ratios and augmentation factors are simultaneous. In addition, higher rates of expenditure on innovative effort—*R&D* lab funding, for example—shift out the frontier and are assumed to be determined by the relative cost of innovative effort and the total costs of production; specifically, expenditure increases with production costs and decreases with the cost of effort.

Before describing the model further, a few additional words are in order about the satellite technology it is intended to represent (more detail is in my 1986 paper). Along with spectrum, satellites use hardware inputs —the spacecraft itself, and the antennae, electronics, batteries, and solar cells that comprise the communications and power-generating subsystems. Spectrum and hardware are substitutable—the principal tradeoff for a given level of communications output involves either generous spectrum use or a more hardware-intensive configuration of communications and power subsystems.

With this production relationship in mind, assume that firms choose cost-minimizing levels of hardware (H) and innovative effort (E) to produce output (Q) subject to an exogenous hardware factor price (P_H), cost of innovation effort (P_E), and rationed quantity of spectrum (S^R). Factor augmentation for a unit rate of innovation effort is represented by input coefficients ($A, g(A)$) which describe tradeoffs along an innovation possibilities frontier described by $g(A)$. Changes in the rate of effort are assumed to shift the frontier homothetically according to the productivity of effort ($h(E)$, with $h'>0$, $h''<0$). Firms thus act as follows:

$$(1)\qquad \min_{H,E,A} P_H H + P_E E$$

subject to

$$\text{(a)}\quad Q = Q(Ah(E)H,\ g(A)h(E)S^R);$$

$$\text{(b)}\quad S \leq S^R.$$

Using α to denote the shadow value of spectrum, and Q_1 and Q_2 the marginal products of H and S, respectively, the first-order conditions can be combined to obtain

$$(2)\quad \alpha/P_H = (g(A)Q_2)/(AQ_1)$$

$$(3)\quad (h'(E)/h(E)) = P_E/(P_H H + \alpha S^R)$$

$$(4)\quad g'(A) = -(HQ_1)/(S^R Q_2).$$

By (2), the ratio of input prices is equal to the ratio of augmented factor marginal productivities. In (3), relative changes in the rate

of innovative effort increase with its cost and decrease with total factor costs; together, (3) and $h'' < 0$ also imply that E increases with total costs and decreases with P_E. By (4), the slope of the innovation possibilities frontier equals the slope of the isoquant for a given H/S ratio.

For a given level of E and factor prices, the simultaneous solution to (2) and (4) determines factor augmentation and input ratios. In the presence of an input price distortion (such as that imparted by spectrum regulation) this joint solution thus indicates the direction of bias induced in factor augmentation and input use. Biases in the rate of innovation can be inferred from (3).

II. Implications for Technical Change

The information required to evaluate the model are data on spectrum shadow values, hardware prices, and the natures of the production function and the innovation possibilities frontier. (These data are described only briefly below; more detail is in my earlier paper.)

Estimates of spectrum shadow values indicate wide variation across users; while this observation is not surprising, what is of special interest here is that the shadow value of spectrum to the satellite industry is some 80 percent larger than analogous values for next-best users. The discrepancy primarily reflects differences in the engineering efficiency of satellite spectrum use. Satellites have a geographically wide broadcasting capability and subsequently distance-independent cost functions, in contrast to alternative, next-best terrestrial technologies requiring a web-like network of land lines for which cost increases with distance.

For estimating the production technology, engineering data are available which have the advantage of being unconstrained by regulatorily mandated spectrum allocations. Econometric estimation of these data suggests that a modified variable elasticity of substitution (*VES*) production function allowing the elasticity of substitution to increase monotonically with H/S best describes satellite technology.

Because much less information is available concerning the nature of the *R&D* process, the innovation possibilities frontier is assumed to have the concave shape typically described in the literature. Results below allow for different shapes of a concave frontier, however, ranging from a symmetric quarter circle to elliptical surfaces with varying eccentricities, to approximate different rates of change in the opportunity costs of factor augmentation.

Turning now to the evaluation of (2) and (4), Figure 1 presents a graphical solution for the joint determination of H/S and $g(A)/A$ as α/P_H increases, given parameters of the production function and innovation frontier. Degrees of eccentricity of the frontier are denoted by β; $\beta = 1$ represents a quarter circle and $\beta < 1$ and $\beta > 1$ describe ellipses that are oblate, respectively, toward the A axis (such that the opportunity cost of spectrum augmentation increases more rapidly compared with the tradeoff implied by the quarter circle) and $g(A)$ axis (where opportunity cost increases less rapidly).

From Figure 1, both H/S and $g(A)/A$ increase with α/P_H. In addition, while opportunity costs represented by differently shaped (although concave) innovation menus matter, they matter inconsistently for ameliorating the distortions. That is, the upward bias in $g(A)/A$ increases (decreases) while the upward bias in H/S decreases (increases) with $\beta < 1$ ($\beta > 1$). Consequently, inefficient spectrum allocations that bias upward the shadow value of spectrum, as suggested by the shadow value estimates, also bias upward H/S and $g(A)/A$, and one or the other distortion persists with certain changes in the nature of the innovation frontier.

Figure 1 also suggests rough magnitudes of the regulatory bias. Points along each function represent a range of values of α/P_H, increasing from .05 to .5 (in increments of .05). Using a mean value of P_H of \$6.2 million (1980 dollars) and assuming (from my earlier paper) that $\alpha =$ \$2.3 million under inefficient spectrum allocation and would be between \$.9 and \$.5 million when efficiently allocated (the shadow value estimates imply

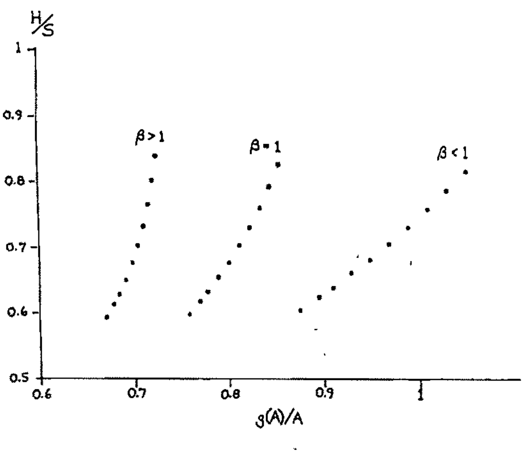


FIGURE 1. SOLUTIONS TO (2) AND (4) AS α/P_H INCREASES

that satellites would have to "outbid" next-best spectrum users by this amount), H/S is 20 to 27 percent larger and $g(A)/A$ is 7 to 10 percent larger in the presence of inefficient regulation (along $\beta = 1$).

To indicate the significance of distortions in $g(A)/A$, the compounding effect that biased induced technical change exerts on distorted input ratios can also be measured. When the model in (1) is estimated *without* taking innovation possibilities into account, the upward bias in H/S is approximately 14 percent. Consequently, biases in technical change can be responsible for a large increase (on the order of 40 to 90 percent (with $\beta = 1$)) in the distortion of factor ratios.

Finally, using H/S and a range of H and S values identified by the engineering data, total costs of production are higher (on the order of 25 percent) with the spectrum constraint. Hence it can be shown from (3) that the rate of innovation effort is also higher under the constraint.

III. Out of Space?

These results must be qualified in several important respects. First, the model assumes that firms engage in technical change to reduce current costs at given factor prices, whereas technical change is in fact more likely governed by expected cost minimization. Satellites are characterized by long lags between the date a spacecraft construction contract is signed and the date the satellite is launched and operating. Although to some extent this delay reflects a regulatory lag itself attributable to delays arising from centralized spectrum management, price expectations probably do play a large role in the implementation of new satellite technology.

Second, spectrum values that might prevail under more efficient allocation are difficult to pinpoint. Spectrum is also used for military and other public sector activities (such as weather data collection), and for private sector services in addition to those modeled here. While the sizes of regulatory biases reported above are rough estimates at best, recall, however, that the relative distortions in Figure 1 require only that spectrum shadow values differ across users.

Third, conclusions concerning whether biases in induced innovation operate in the same direction as biases in input ratios are sensitive to the structure and parameters of the production function and innovation possibilities frontier. Increases in the relative price of an input will not necessarily bias technical change toward augmentation of that input because movement along the innovation frontier depends on factor shares. For instance, in contrast with results reported here, a *CES* technology and quarter-circle menu yield distortions in directions that depend on the substitution parameter (see Okuguchi).

Subject to these caveats, this research suggests that more efficient spectrum allocation, rather than exhortations for greater innovation targeted towards spectrum use, should be the goal of public policy toward the satellite industry. Certainly these caveats themselves underscore how opportunities for technical change exacerbate the information-intensive requirements for spectrum regulation. More flexible allocation schemes (as provided by spectrum markets or lotteries permitting resale—reforms already advocated in the existing economics literature) may do a better job in signalling the pace and direction of innovation.

Also indicted is public spending on satellite *R&D* to augment spectrum. The NASA project cited earlier has a budget of $500 million (a large expenditure for a single NASA *R&D* program). At this level, the project has a ratio of spending to new spectrum regions the research is expected to exploit of about $7.2 million per signal, a value some 8 to 14 times larger than estimates of marginal spectrum shadow values. To be sure, the program is intended to develop a variety of new technologies, some of which are less directly related to spectrum. Nonetheless, it may pay to reassess the incremental benefits of public spending on spectrum *R&D*, as well as redress the adverse consequences of spectrum regulation for other public sector space programs.

REFERENCES

Besen, Stanley M. et al., *Misregulating Television*, Chicago: University of Chicago Press, 1984.

Kamien, M. I. and Schwartz, N. L., "Induced Factor Augmenting Technical Progress from a Microeconomic Viewpoint," *Econometrica*, October 1969, *37*, 668-84.

Levin, Harvey J., *The Invisible Resource*, Resources for the Future, Baltimore: Johns Hopkins University Press, 1971.

Macauley, Molly K., "Regulation and Technical Change in Communication Satellites," Discussion Paper EM86-01, Resources for the Future, January 1986.

______ **and Portney, Paul R.,** "Property Rights in Orbit," *Regulation*, July/August 1984.

Magat, Wesley, "Regulation and the Rate and Direction of Induced Technical Change," *Bell Journal of Economics*, Autumn 1976, *7*, 478–96.

Okuguchi, Koji, "The Implications of Regulation for Induced Technical Change: Comment," *Bell Journal of Economics*, Autumn 1975, *6*, 703–05.

Smith, V. Kerry, "The Implications of Regulations for Induced Technical Change: Reply," *Bell Journal of Economics*, Autumn 1975, *6*, 706–07.

______, "The Implications of Regulation for Induced Technical Change," *Bell Journal of Economics*, Autumn 1974, *5*, 623–32.

Federal Communications Commission, Space WARC Advisory Committee, *Second Advisory Committee Report*, January 31, 1985, 31–34.

Federal Register, Vol. 48, September 6, 1983, 40233–260.

National Aeronautics and Space Administration, "Advanced Communications Technology Satellite (ACTS): Notice of Intent for Experiments," Washington, D.C., November 1984.

U.S. Congress, House, Subcommittee on Space Science and Applications, *Hearings on the 1985 NASA Authorizations*, 98 Cong., 2d sess., Washington, 1984.

______, Office of Technology Assessment, *Radiofrequency Use and Management: Impacts from the World Administrative Radio Conference of 1979*, Washington: USGPO, 1982.

SITING OF HAZARDOUS FACILITIES†

Property Rights, Protest, and the Siting of Hazardous Waste Facilities

By ROBERT CAMERON MITCHELL AND RICHARD T. CARSON*

In 1980, the U.S. Environmental Protection Agency estimated that between 50 and 125 new sites for hazardous waste facilities (*HWFs*)[1] would be needed in the near future. Since that time no major *HWF* has been sited anywhere in the United States. The EPA had anticipated that local opposition would make finding these sites an "exceptionally difficult task." Their pessimism was well founded and, if anything, understated. According to the Hazardous Waste Counsultant's latest state-by-state review, the outlook for siting *HWFs* in the future is "even more bleak" than in the past, due in large part to what they term a worsening of the "emotional atmosphere" surrounding siting efforts. This failure to site any new *HWF* has come about in spite of assurances by government and company officials that new facilities built according to the present standards would pose negligible risks to the local residents. Attempts have been made to break the deadlock by instituting extensive public participation procedures, establishing state siting boards with the power to overrule local decision makers, and requiring facility owners to compensate local governments for safety services the latter provided.

In this paper, we argue that the ambiguous nature of the present property rights governing the siting of *HWFs* is an important cause of the stalemate. We offer a new approach to siting which recognizes the de facto property rights assumed by local communities. We propose a political market, via a referendum mechanism, for allocating *HWFs*. The referendum, supervised by the state, would be held at the request of the firm wishing to site the *HWF* with the developer, in effect, offering a comprehensive package of incentives for the community in exchange for a yes vote.

To understand the rationale for our approach, it is first necessary to examine the evolving nature of the property rights for siting a *HWF*. The driving forces are changing perceptions of the risks associated with toxic waste disposal and a social movement of considerable power which has raised the cry of "not in my backyard." We show that rational citizens have much to gain by opposing the siting of new hazardous waste facilities near them. Their resistance, however, imposes large costs on society as a whole, since as quantities of toxic chemicals are being held in temporary and deteriorating storage conditions as they await destruction, or a permanent home, strong incentives are created for illegal "midnight" dumping.

I. The Problem

Hazardous wastes are a by-product of the chemical revolution that followed World War II. Until recently, disposing of wastes was not considered to be a social problem and dumps with hazardous materials in them were treated by the public and planners as minor extensions of their garbage dump and sanitary landfill cousins. Opposition, if any, was based on their nuisance characteristics, not

†*Discussants*: Allen V. Kneese, Resources for the Future; William D. Schulze, University of Colorado.

*Senior Fellow, Resources for the Future, 1616 P Street, NW, Washington, D.C. 20036; and Assistant Professor, Department of Economics, University of California, San Diego, La Jolla, CA 92093, respectively.
We thank Peter Navarro, Susan Pharr, Paul Portney, and Clifford Russell for helpful comments.

[1]These include waste treatment facilities, landfills, and incinerators.

on perceived safety risks. The property rights status quo was one in which the developer's entitlement to engage in waste handling activities was preeminent as long as the facility was located in an industrial area.

The passage of the Resource Conservation and Recovery Act (RCRA) by Congress in 1976 marked official recognition that these wastes, many of them disposed of improperly in the past, posed a potentially serious health threat. However, widespread public awareness of possible danger to local communities from this source did not take shape until two years later when the problems at Love Canal reached the national news media. Congress passed the Superfund legislation in 1979 to clean up existing toxic waste dumps. The entire town of Times Beach, Missouri, was abandoned after finding dioxin contamination in 1982, and news reports of contaminated drinking water wells have now become commonplace. Proposed *HWFs* quickly became the subject of widespread protest despite the fact that they had to meet the stringent federal design and operation safety standards imposed by RCRA and further augmented by additional state regulations. The Sand Canyon facility in Los Angeles County illustrates this situation. Four years of work and $1.5 million were spent on a proposed comprehensive treatment and land disposal facility before its corporate owner withdrew the proposal in the face of seemingly insurmountable public opposition. In Texas, a regional authority proposed a high temperature incinerator for toxic wastes from the area (a solution favored by environmentalists). Despite a well-demonstrated need for such a facility and initial support from surrounding local governments, citizen opposition caused the developer to withdraw the proposal after a 3-year battle when it became apparent that political approval was not going to be forthcoming.

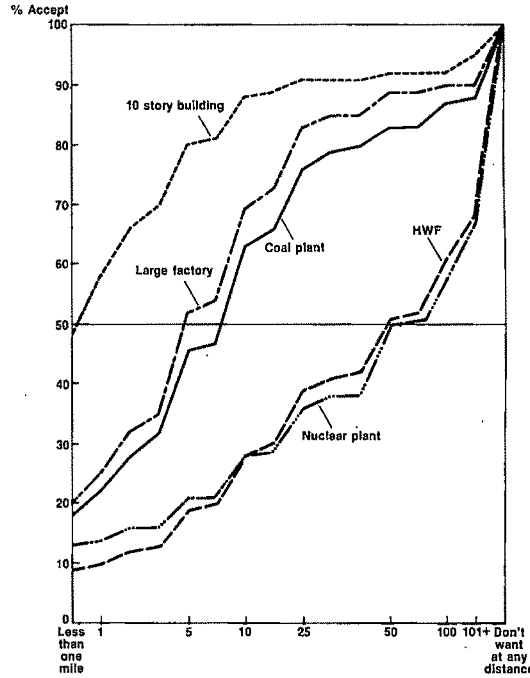


FIGURE 1. CUMULATIVE PERCENTAGE OF PEOPLE WILLING TO ACCEPT NEW FACILITIES AT VARIOUS DISTANCES FROM THEIR HOMES

II. Aversion Profiles

The NIMBY acronym (not-in-my-backyard) aptly captures the views of those who resist facility siting. The syndrome itself is not new—homeowners have long resisted the siting in their neighborhood of undesirable facilities. What is new is the scale and intensity of protests provoked by facilities perceived to be risky. The present high level of risk averseness is illustrated in Figure 1 which shows the percentage of the public in a national survey[2] willing to accept (without protesting or moving) each of five hypothetical facilities as a function of the distance of the facility from their residence.

Three distinct "siting aversion profiles" emerge, with corresponding "backyards" and protest constituencies. Reactions to a 10-story office building represent a useful baseline.

[2]These data are from a survey conducted by Resources for the Future (Mitchell, 1980). The general shape of the profiles has been found to be robust against alternative question wordings and the addition of other types of facilities.

Over half say they would accept such an edifice if it were at least a mile from their houses. Majority acceptance of either the industrial plant or the coal-fired electric power plant, facilities that are likely to be perceived as dirty and potentially obnoxious neighbors, occurs at approximately 9 miles. In contrast, the two facilities posing potentially catastropic but extremely low probability risks, a nuclear power plant and a new, well-regulated disposal site for hazardous wastes, reach majority acceptance only at the 50-mile mark, a distance "premium" of 49 miles from our arbitrarily selected baseline. This suggests a crucial difference between the siting of an ordinary industrial facility and a *HWF*: the "neighbors" affected by the latter involve entire communities. Another difference is the number of people who feel strongly about the issue. Whereas only 9 percent expressed the extreme view that they did not want the two industrial facilities as neighbors "at any distance," 29 percent expressed such a view about the two "risky" facilities.

III. Protest Mobilization

At the local level, the aversion to *HWF*s is translated into active protest whenever new facilities are proposed. Why do local residents protest? Mobilization is facilitated by: 1) the high cost perceived to be imposed on the local community by the *HWF*, 2) the low cost of protesting, and 3) the high probability of success.

First, *HWF*s are a prime example of a regulated entity whose costs and benefits are so distributed that the former are concentrated, while the latter are distributed, far beyond the local area. The principal costs believed to be posed by *HWF* are the health risks posed by groundwater and soil contamination in the case of landfills, and contamination of the air by cancer-causing substances in the case of incineration facilities. The high level of perceived risks may be attributed both to the institutional context in which these risks occur and to the nature of the risks themselves.

The news media have highlighted past failures to handle toxic wastes properly and scientific uncertainties about the risks they pose to the public. At the local level, the siting issue appears as an abrupt threat that involves a visible source (the site) for which clear responsibility can be ascribed (the developer)—characteristics that heighten public awareness of the perceived risk. In contrast to nuclear power plants or industrial plants, for which there is usually a local constituency, a *HWF* provides few benefits such as jobs or tax revenues (A. D. Tarlock, 1984). Finally, residents may fear the decline of local property values.

The degree of concern about the risk externality posed by *HWF*s is strongly influenced by the nature of the perceived risk. The risks posed by these facilities include characteristics which have been shown in other contexts to be strongly associated with risk aversion (Paul Slovic et al., 1980). They are perceived as: 1) involuntary (imposed on the community without its consent); 2) lethal; 3) memorable (due to being subject to arresting media coverage); 4) not susceptible to personal control; 5) persistent (having the potential to effect future generations); and 6) unfair (since most of the benefits accrue to those living far beyond the geographic area subject to risk).

Two characteristics of siting controversies help lower mobilization costs. First, the local character of the controversy makes it easy to identify and communicate with potential protesters. Geographic concentration also allows use of preexisting social networks and institutions (such as churches and neighborhood organizations) for leader and member recruitment purposes. This reduces organizational costs and makes free riding easier to manage through informal social control in the form of pressure to participate. Second, public participation procedures used in many siting processes, such as hearings, offer a focal point both for organizing and for news media coverage, and easy access to decision makers.

For individual participants, the cost of mobilization involves time and money. This includes time spent in activities such as recruitment, fund raising, and organizational maintenance, as well as time spent in protest activities such as writing letters, working on

lawsuits, and organizing and attending rallies. The time commitments necessary for a successful protest movement are lumpy: only a relatively small number of activists need to commit substantial amounts of time to the effort. For most participants, only occasional participation is necessary, because much is demanded of only a few.

The third factor affecting mobilization is the perceived likelihood that the protest activity will benefit the participant. Some people, usually highly committed activists, derive utility from the act of protest itself, which confirms their values and sense of self-worth. The efficacy calculus for ordinary participants normally involves a belief that their cause has some chance of achieving its goals. Factors that contribute to a sense of efficacy in siting protests include the widespread support for the protest in the affected community, the frequent sympathy or even support for the protest on the part of local elected officials, the availability of proven tactics (ranging from sit-ins and demonstrations to lobbying and legal interventions), expertise (from national organizations), and arenas in which to contest and delay the siting (such as local hearings, the courts, and, of particular importance, local zoning and permiting processes).

IV. Evolving Property Rights

Property rights specify how persons may benefit or be harmed and, therefore, who must pay whom to modify the actions taken by affected parties. In a now famous article (1960), Ronald Coase argued that the assignment of property rights to one party or another did not, in the absence of transactions cost, affect economic efficiency, although it did affect the distribution of wealth. Coase's insight was a deep one: resources would be put to their most efficient use regardless of how the political system initially chose to allocate property rights. The problem with the hazardous waste situation is that currently no one really has clear title to site a *HWF*—not the firm, not the community, and not the residents as individuals.

Harold Demsetz correctly saw that property rights were subject to change over time to "accommodate externalities associated with important changes in technology or market values" (1967, p. 350). Firms wishing to site a *HWF* lost their unfettered right to locate where they wished as the public and government officials became alarmed over the possible risks posed by the technology. Local residents have obtained increasing ability to delay (and thus effectively block) siting efforts in adminstrative and judicial hearings. Local communities have taken a leading role in stopping the construction of new *HWF* through the use of their extensive police powers to regulate zoning and safety matters. With a few exceptions, however, communities do not have the legal right to ask for sizeable payments in exchange for issuing the necessary licenses and permits.

The recent establishment of state siting boards with the power to preempt local governments represents an attempt to reassert the former property right regime. The concurrent establishment of schemes for compensating communities for the presence of a *HWF* represents a movement in the opposite direction—toward giving the property right to the community. The innovative Massachusetts' siting law (M. O'Hare et al., 1983) has both features, going further in the direction of bargaining for compensation and less in the direction of preemption (calling for binding arbitration only in the case of irreconcilable differences) than any law in the country. No facilities yet have been sited under this law, suggesting that compensation without ultimate local veto power over a facility may not be a successful strategy. But if local residents were *individually* to hold the property right, developers could not bargain efficiently with the large number of potentially affected residents and one holdout could block a well-conceived project.

V. Community Property Rights: A Proposal

One possible solution is to recognize a collective property right by having states pass a law specifying the use of referenda to determine local approval or rejection of a proposed *HWF* facility. Such a law would require the relevant political authorities to hold a referendum when requested by a qualified

developer meeting state requirements. The terms of the arrangement would be proposed by the developer and incorporated into the ballot proposal. Both the developer and the state, to the extent that it desired the siting, would have strong incentives to develop winning proposals. Developers obviously would aim at selecting potential sites where voters would be more likely to agree to the least expensive package of measures designed to compensate a community for accepting the *HWF*. Designing the package and promoting it would necessarily involve the equivalent of a public participation program. Naturally the costs of the package would be passed on to enterprises that wished to use the facility. In order for such a proposal to be viable, there would have to be enough technically acceptable sites available so that the political market could be sustained, and no single community would have a siting monopoly.

A large number of possible compensatory measures have been suggested in recent years. The contents of a developer's particular package could vary according to the nature of the facility, the characteristics of the site, and the community's concerns. The types of measures which might be included are: guarantees against property value declines, incentive payments to the community (which could be earmarked to reduce property taxes or for other purposes), outside monitoring,[3] accident insurance, credible guarantees of nonabandonment, donation of land for use as parks, and in-kind services such as free waste disposal for community residents and businesses.

Should the decision rule be a simple majority, or something larger, such as the often used two-thirds majority? Although a two-thirds majority requires a more expensive package, we argue that it is more likely to result in a Pareto-improving outcome and greater community harmony. Who would administer and enforce and contract established by the referendum? This would undoubtedly fall to the local political authorities first, and ultimately to the state. Doubts about enforcement would only increase the payments required to pass the referendum. There must be sufficient administrative flexibility to respond to new EPA regulations and to technological change. How should the boundaries defining who should be allowed to vote on the proposal be defined? This is an admittedly difficult political question which the state legislature would have to decide.

Assigning the right to refuse a risk externality to those who claim it, and exercising coercion only to the extent of requiring them to vote on legitimate offers to compensate them for accepting the risk, has several desirable properties in this case. The developer and the state have strong incentives to address the issues of most concern to the community, and the state's role is more consistent with its interest in the outcome. The community's incentive to be intransigent is minimized because it has the power to say no. The community is presumably protected from unwittingly accepting too great a risk because the facility would have to meet strict federal and state safety regulations. Moreover, the debate occasioned by the referendum should ensure close scrutiny of the developer's proposal. Paying for the compensation package transforms the hitherto concentrated costs on the local community into more equitably shared burdens that are borne by the ultimate beneficiaries of the facility. Finally, to the extent that this increases the costs of handling hazardous wastes, those who produce the wastes will have an incentive to engage in in-plant waste-stream modifications and resource recovery.

[3] If the developer or government is not trusted by the community to monitor the facility, the cost of a winning compensation package may be drastically increased. Monitoring by an outside agent, such as an environmental organization, might reduce the cost of the package's other elements.

REFERENCES

Coase, R. H., "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

Demsetz, H., "Toward a Theory of Property Rights," *American Economic Review Pro-*

Asymmetries in the Valuation of Risk and the Siting of Hazardous Waste Disposal Facilities

By V. KERRY SMITH AND WILLIAM H. DESVOUSGES*

Recently several economists (Richard Thaler, 1980; Jack Knetsch and J. A. Sinden, 1984), following suggestions of psychologists (Daniel Kahneman and Amos Tversky, 1979), have argued that current economic models of consumer behavior fail to explain observed asymmetries in how individuals respond to gain vs. losses in perceived entitlements. Attention to their arguments is increasing because they relate to many current policy issues—especially those associated with undesirable land uses.

Most of the papers suggesting this limitation with the conventional economic framework have been motivated by the large differences between the estimates of willingness to pay vs. willingness to accept as measures of the change in individual well-being that would result from a change in the conditions of access to (or the quality of) a commodity.[1] In this paper we report the first evidence of a sizable property rights effect using only willingness-to-pay measures. This change is potentially important because both the recent appraisal of contingent valuation surveys (see Ronald Cummings et al., 1986), an important source of the available empirical evidence, and laboratory experiments suggest that individuals may have difficulty in dealing with the concept of compensation. This is especially true when there is no opportunity for individuals to learn about transactions that involve compensation through experience.

Based on a contingent valuation survey of households in suburban Boston, we found that respondents bid significantly more to reduce risk than they indicated they were willing to pay to avoid an equivalent risk increase. While our findings support suggestions that changes in the implied entitlements (to safety) can lead to large differences in welfare measures for risk changes, several of these earlier arguments would have implied that individuals were willing to pay more to avoid a risk increase—the opposite to our results. Thus, these differences imply that the determinants of individuals' valuations for risk changes are more complex than past studies have acknowledged.

I. Defining and Describing Individual Values for Risk Changes

Our measure of an individual's valuation of a risk reduction is an option price—a constant, state-independent payment for a reduction in the probability of premature death. While often described in the literature (see, for example, Michael Jones-Lee, 1974) as the individual's willingness to pay, the option price is actually one of an infinite number of possible dollar measures of the change in expected utility associated with a change in risk. The other concepts would describe an individual's valuation under different prospects for contracts with state-dependent payments (Daniel Graham, 1981).

Our empirical estimates are based on a survey of 609 households in suburban Boston. To elicit their valuation of changes in the risk of exposure to hazardous wastes, we explained the equivalent of a two-stage lot-

*Centennial Professor of Economics, Vanderbilt University, Nashville, TN 37235, and Senior Economist, Research Triangle Institute, respectively. Thanks are due Howard Kunreuther for comments on an earlier draft of this paper. Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency through Cooperative Agreement CR-811075, it has not been subject to the Agency's peer and administrative review and therefore does not reflect the view of the Agency. No official endorsement should be inferred.

[1] Conditions of access refers to the terms governing with the exchange of a commodity (i.e., rationing exclusively by a single price, multiple prices, or nonprice conditions, and the timing of the payments).

tery with two outcomes—life and death. Respondents were told that the first stage involved a chance of being exposed to hazardous waste, described as a probability R. Given they were exposed, the second stage posed a separate risk, q, (explained as based on their health and heredity) of premature death resulting from the exposure (i.e., a conditional probability). Government regulations on the operations of a nearby (hypothetical) landfill receiving hazardous wastes were described as the source of the control of their exposure risk, R.

Instead of an annual risk for each component of the lottery, we described the temporal dimension of the risk as exposure sufficient to provide the prospect (Rq) of premature death in 30 years.[2] This description reflects the physical processes associated with these risks. To explain the baseline risks, we used circles with darkened slices along with ratios and percentages for all three probabilities—R, q, and Rq. Respondents were given several different cards to depict the proposed changes in these risks.

To examine the effects of the implicit property rights (or perceived entitlements) to risk, we presented the same risk change to each respondent in two different ways. First, we described regulations that would reduce exposure risk and asked for the maximum (monthly) payment to obtain each of two proposed reductions. Later in the questionnaire, we described the second situation in which an increase in the hazardous wastes disposed in a hypothetical landfill near their homes would increase their risk. Additional regulations could be imposed to avoid the risk increase. This question asks each respondent's willingness to pay to have these regulations and thereby avoid the risk increase. The same ending risk and risk change were used in the two types of questions.

Trained interviewers collected our data with in-person interviews from approximately 85 percent of the enumerated sample households. Our survey design varied the baseline exposure risk R, the conditional probability q, and the level of government proposed as the source of the decision to allow the risk increase in the second valuation question.[3] The first proposed reduction in R was 50 percent of the baseline value and the second 40 percent of the first reduction's endpoint. The design randomly assigned questionnaires to our interviewers and respondents.[4]

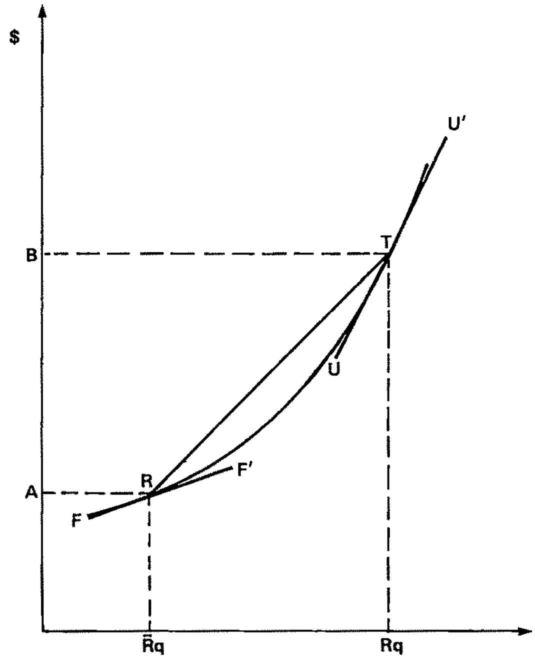


FIGURE 1. MARGINAL VALUATION OF RISK CHANGES

[2]To use estimates of annual risks would ignore the results of the theoretical literature suggesting that temporal risks can lead to violations of the independence axiom of the expected utility framework when an individual can adjust to the uncertainty (see Mark Machina, 1984).

[3]There were 8 design points in our survey, allowing for differences in R and q. Each of three vectors of high values for R were paired with two values for q. In addition, 2 low-probability design points were considered. The three sets of R changes were (1/5 to 1/10, 1/10 to 1/25), (1/10 to 1/20, 1/20 to 1/50), and (1/30 to 1/60, 1/60 to 1/120). These were each matched with values of 1/10 and 1/20 for q to define 6 design points. In addition, we considered one set of low-risk values for the exposure risk—(1/300 to 1/600, 1/600 to 1/1500) paired with each of two values for q—1/100, 1/200.

[4]The survey had several objectives. Sixty percent of the sample received the direct valuation questions and are the focus of our analysis.

Using the simple formulation of the expected utility model, Figure 1 illustrates why our two questions should provide the same measure for the incremental value of a risk change. The term $y(Rq)$ describes the individual's valuation of risk change. It is the expenditures required to maintain a given level of expected utility for alternative values of Rq. The usual economic analysis of risk changes would imply that $y(\cdot)$ should be steeper at Rq than $\bar{R}q$ (where $\bar{R}q < Rq$).[5] However, our questions elicit the equivalent of AB, the individual's bid for a change from Rq to $\bar{R}q$, or the maximum payment to avoid the reverse. Both should yield the same estimate of the incremental value of $(Rq - \bar{R}q)$ even though the slope of $y(\cdot)$ differs at these two points (as illustrated by FF' vs. UU' in the figure).

II. Entitlements and the Valuation of Risk Changes

To estimate these models for evaluating the effects of entitlements we used James Heckman's (1979) two-step approach using the estimated Mills' ratio to correct for the selectivity effects arising from restricting the sample to positive bids. Table 1 presents the ordinary least squares estimates for three models. Each model is a semilog function with the logarithm of the individual's reported value for a risk change scaled by the magnitude of the risk change (i.e., an estimate of the marginal value of risk based on the valuation response) specified to be a function of the determinants of the baseline risk (R and q) and a set of variables to control for the socioeconomic characteristics and attitudes of each respondent. In each case our first-step model hypothesized that nonzero bids reflected greater understanding of the terms of the contingent valuation questions. Thus each probit model was specified to be a function of both features of the question and of the individual that might be associated with his understanding the terms of the hypothetical market. These features include: R, q, education, knowledge of hazardous waste problems (measured by number of news articles on hazardous waste that were reported to have been read in the past 3 months), an index of the individual's feelings concerning the effectiveness of the

TABLE 1—MODELS FOR INCREMENTAL OPTION PRICE RESPONSES: SEMILOG MODEL[a]

Independent Variables	Risk Increase	Risk Decrease	Pooled Sample
I	−2.57	−2.86	−2.35
	(−4.27)	(−5.06)	(−5.30)
R	−.03	−.11	−.04
	(−3.98)	(−3.58)	(−8.04)
R^2	$.07\times10^{-3}$	$.79\times10^{-3}$	$.13\times10^{-3}$
	(2.35)	(1.26)	(5.77)
q	−.02	−.01	−.02
	(−5.62)	(−3.67)	(−5.77)
y	.02	.02	.02
	(4.30)	(4.28)	(6.03)
a	−.01	−.01	−.01
	(−1.21)	(−2.53)	(−2.57)
h	.07	.05	.07
	(2.76)	(2.48)	(3.91)
c	.04	−.07	−.03
	(0.77)	(−1.19)	(−0.59)
v	2.31	2.84	2.95
	(5.06)	(6.61)	(9.72)
P	–	–	−1.50
			(−8.08)
G	–	.18	.16
		(1.30)	(1.16)
λ	−.40	−.69	−.63
	(−0.80)	(−1.52)	(−1.79)
r^2	.85	.77	.79
ρ	.13	.30	.25

[a]The values shown in parentheses are t-ratios using Heckman's corrected covariance matrix. The variable definitions are as follows: I = intercept; R = baseline risk of exposure to hazardous waste; q = baseline conditional probability of death (each scaled by 1000); y = household income in thousands of dollars; a = age in years; h = index (from 1 to 10) indicating perceived seriousness of hazardous waste pollution (10 = most harmful); c = number of related children in households; v = qualitative variable for low probability design points (see fn. 4); P = qualitative variable = 1 for questions concerning avoiding a risk increase; G = qualitative variable for source of proposed increase 1 = government decision, 0 = town council decision; r^2 = R squared for the equation: λ = inverse Mills ratio; ρ = squared correlation between probit and selection model disturbances.

[5]Given concave state dependent utility functions and the assumptions that both total utility and marginal utility of income were higher in the state living, the marginal value will decline with decreases in the baseline risk. See Jones-Lee or Milton Weinstein et al. (1980).

local water district, and residence in Acton (a town in suburban Boston that experienced several contamination incidents in the recent past). The third column which reports the results for a model estimated with the bids for risk reductions pooled with those for avoiding risk increases is the most relevant for judging the effects of entitlements. These effects are reflected in two variables: P is a qualitative variable ($=1$ for avoiding risk increase, 0 otherwise) that distinguishes responses to risk reductions from those for avoiding a risk increase; and G relates to one element of the context for the risk-avoidance question—the increase in hazardous wastes was described as approved by either the government ($G=1$) or the individual's town council ($G=0$).

Our estimates suggest that the marginal values implied by the bids to avoid a risk increase are significantly smaller than those to obtain an equivalent risk as a reduction. There is a tenfold difference in the implied marginal valuation of risk. This finding is inconsistent with the direction of the effects of entitlements described by Thaler, Knetsch-Sinden, and Kahneman-Tversky. Their arguments have maintained individuals are willing to pay *more* to retain a desirable commodity (or a low level of risk) than to purchase it. Our findings indicate just the opposite conclusion. However, there is an important caveat. We believe that these responses reflect an indirect response to the endowment effect of the risk avoidance question. The respondents felt they should *not have to pay to maintain their stated low level of risk*.

While one might suggest that the bids for avoiding risk increases compared to those for risk reductions reflect a diminishing marginal valuation of risk (with decreases in the risk level), this is not consistent with other aspects of our results. That is, the signs for the coefficients for R and q in the third column along with our explanation accompanying Figure 1 cast doubt on this argument. (Cols. 1 and 2 indicate similar results.)

We believe that the explanation for these results lies in how individuals perceive entitlements that extend beyond simple titles to marketable commodities. Since applied welfare measures have relied on the individual's budget constraint to model the influence of property rights on his ability to realize any particular utility level, conclusions based on this framework do not reflect the individual's perceived rights to low risks or to other nonmarketed commodities. In much the same way as tradition, common law, or accepted practices can lead to capitalization effects comparable to formal property rights, perceived entitlements may well affect how individuals will report values for involuntary losses of implicit rights.

REFERENCES

Cummings, Ronald G., Brookshire, David S. and Schulze, William, *Valuing Public Goods: The Contingent Valuation Method*, Totowa: Rowman & Allanheld, 1986.

Graham, Daniel A., "Cost-Benefit Analysis under Uncertainty," *American Economic Review*, September 1981, *71*, 715–25.

Heckman, James J., "Sample Biases as a Specification Error," *Econometrica*, September 1979, *47*, 153–62.

Jones-Lee, Michael, "The Value of Changes in the Probability of Death or Injury," *Journal of Political Economy*, July/August, 1974, *82*, 835–50.

Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decisions under Risk," *Econometrica*, March 1979, *47*, 263–91.

Knetsch, Jack L. and Sinden, J. A., "Willingness to Pay and Compensation Demanded: Experimental Evidence of an Unexpected Disparity in Measures of Value," *Quarterly Journal of Economics*, August 1984, *99*, 507–21.

Machina, Mark J., "Temporal Risk and the Nature of Induced Preferences," *Journal of Economic Theory*, August 1984, *33*, 199–231.

Thaler, Richard, "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organization*, March 1980, *1*, 39–60.

Weinstein, Milton C., Shephard, Donald S. and Pliskin, Joseph S., "The Economic Value of Changing Mortality Probabilities: A Decision-Theoretic Approach," *Quarterly Journal of Economics*, March 1980, *94*, 373–96.

A Sealed-Bid Auction Mechanism for Siting Noxious Facilities

By HOWARD KUNREUTHER AND PAUL R. KLEINDORFER*

This paper proposes a sealed-bid mechanism for facilitating the siting process of noxious facilities such as prisons, trash disposal plants, and incinerators for hazardous waste. These facilities have created such opposition by communities that the acronyms NIMBY (not-in my backyard) or LULU (local undesirable land use) are now in common usage to describe the public's reaction to them.

The noxious facility siting problem arises because there are economies of scale associated with having only one plant to serve the needs of a wide region. The community that hosts the plant absorbs all of the environmental costs, while the rest of the region enjoys the benefits of the facility. One reason that these facilities have been so strongly opposed by communities is that they generate relatively little new employment and provide limited additional taxes in relation to their perceived negative impact (D. Morell and C. Magorian, 1982). Some compensation arrangement is thus needed to share the gains from the winners to the potential loser. This paper discusses a sealed-bid auction mechanism for selecting a location for the facility so that all communities feel they are better off than under the status quo.

I. A Sealed-Bid Mechanism with Compensation

The siting of a noxious facility is a mixture of a public good and a private bad. The positive externalities associated with a facility in community i yield positive value to all the other j communities in the area. Community i, however, receives negative value from hosting the plant. To be more concrete, consider the simplified problem of locating a trash disposal facility in one of N communities in a region where $V_i < 0$ represents the negative value to each community i from having the plant in their backyard, and $V_{ij} > 0$ represents the value to community i of siting the plant in location $j \neq i$. Due to economies of scale, only one plant will be chosen from among the N sites and all communities have agreed that they are willing to participate in this regional cooperative. This implies that the compensation arrangements are sufficiently attractive so that the benefits of participating in the cooperative arrangement are greater than the default option.[1]

One of the most difficult problems associated with developing meaningful compensation arrangements for siting noxious facilities relates to preference revelation. Communities may have an incentive to not only exaggerate their willingness to accept (*WTA*) the facility but also to underreport their willingness to pay (*WTP*) for someone else to assume this burden.[2] At first glance the literature on demand-revealing mechanisms appears to be appropriate for inducing individuals to specify accurate *WTA* and *WTP* values. Specifically, F. Clarke (1971) has developed a procedure that leads individuals to declare their true preferences for a pure public commodity by charging a tax that is dependent, in part, on how their responses affect the final outcome. Upon closer inspection, however, one finds that the Clarke preference-revelation mechanism requires the public commodity to have a positive value to each participant so there is a net surplus after the commodity tax is levied. This will not be the case for the

*Professors of Decision Sciences, University of Pennsylvania, Philadelphia, PA 19104. This research was partially supported by NSF grant no. SES-8212123. We thank Peter Knez, James Richardson, Michael Selman, and Rudy Yaksick for helpful discussions and assistance in designing and administering the controlled laboratory experiments.

[1] We are assuming that any community that decides not to enter this arrangement can be excluded from the use of the facility.

[2] For a detailed discussion of alternative estimation procedures for *WTA* and *WTP*, see Ronald Cummings et al. (1985).

class of problems associated with public bads and hence the truth-telling property of the Clarke mechanism fails to hold for the problems of interest here.[3]

We have developed a sealed-bid auction model for eliciting *WTA* values when a single community is chosen as the "winner" from among N possible candidates (see Kunreuther et al., 1985). Each community i announces a *WTA* value, X_i, that it will receive if it is the winner. The understanding is that if another community is selected as the host site then i will have to pay a tax $t_i = X_i/(N-1)$ to help compensate the other town. A regional siting agency assembles all the data, selects the community with the lowest *WTA* value and uses the tax payments to compensate the host site. By using the criterion $\min X_i$ this procedure is guaranteed to be budget balancing and is likely to create a surplus for the regional agency.[4]

The above procedure, the *low-bid auction*, has several additional appealing features. It is a coalition-free mechanism since each community's *WTP* value is independent of any other community's *WTA* value. Hence two or more communities cannot strategically link their bids in order to extract mutual gain from the procedure. It is also designed to be equitable by making the tax on any one community inversely proportional to the number of towns who are part of the consortium.

What bidding strategy will be utilized by a community in specifying its *WTA*? When each community knows its own preferences but has no information on others, then a max-min bidding strategy is prudent. This is an optimal bid to make if each community is considered to be equally likely to make the lowest bid and all candidates are risk averse.[5]

Under the low-bid auction procedure, the max-min bid is given by

$$(1)\quad X_i^* = ((N-1)/N)\left(\left[\min_{j\neq i} V_{ij}\right] - V_i\right).$$

The only communities who would be willing to participate in the low-bid auction are those where $\min V_{ij} > X_i^*/(N-1)$ under the assumption that a max-min strategy is utilized and the default option yields benefits equal to zero. Concerning efficiency, it can be verified (see Kunreuther et al.) that the low-bid auction leads to an efficient outcome (i.e., a location i maximizing $V_i + \Sigma_{j\neq i} V_{ji}$) under max-min bidding when $V_{ij} = V_{ik}$ for all $j, k \neq i$, that is, community i is indifferent as long as the facility is not sited at location i.

To illustrate, consider a community with $V_i = -1000$ and $V_{ij} = 600$ for all j. If $N = 5$ then $X_i^* = 1280$ so that community i will obtain a net gain of 280 whether or not it is the lowest bid. If $V_{ij} < 250$, then it would prefer not to participate in the consortium unless the default option was a negative value (for example, having to pay more than the current price for trash disposal).

II. Experimental Results

A series of controlled laboratory experiments with $N = 5$ using the low-bid auction examined how close individuals were to max-min solution (Kunreuther et al.). An interactive computer program in Lotus123 enabled individuals to try different values of X_i before making their final choice. A sample value matrix is given in Table 1 with the max-min solution for each participant indicated on the bottom row. Payoffs were in a fictitious currency (francs) with each franc worth 1 cent. All participants received private information on their own V_i and V_{ij}. At the end of each auction the participant with the lowest bid was announced but the winning bid was not revealed. A series of 10 different auctions were actually played.

[3] T. N. Tideman and G. Tullock (1976) point this out in their discussion of the Clarke mechanism.

[4] The tax surplus can be utilized to provide common facilities to all individuals participating in the program (for example, a hospital), or for activities related to the trash facility itself such as monitoring and control of midnight dumping.

[5] When each community is viewed as having an equally likely chance of being selected, then the expected value of any bid is the same. Max-min strategy produces the smallest range of outcomes and hence is optimal for a risk-averse community.

TABLE 1—PAYOFF MATRIX AND MAX-MIN BIDS FOR 5-PARTICIPANT LOW-BID AUCTIONS

Location	Participant 1	2	3	4	5
1	−1000	620	640	650	700
2	600	−1100	640	650	700
3	600	620	−1200	650	700
4	600	620	640	−1300	700
5	600	620	640	650	−1500
Max-Min Value	1280	1376	1472	1560	1760

An analysis of the results reveals that the average absolute percentage deviation from max-min bids is relatively small particularly for later periods. Specifically, in periods 7 through 10 in three experiments approximately three-fourths of the 60 different bids were within 5 percent of the max-min values. It should be noted, however, that there were significant departures from max-min bids in the early periods for inexperienced subjects. Hence the importance of learning. An analysis of the data reveals that for 80 percent of the auctions, ther person who elicited the lowest bid was the same one predicted by the max-min strategy.

Our objective is eventually to utilize this procedure in a real world field setting. In this spirit David Brookshire et al. (1985) conducted a field survey related to the delivery of goods by one individual to a number of others who belonged to a cooperative. This situation is conceptually analogous to the noxious facility siting problem since one participant incurs all the costs of delivery while the other members benefit from this service.

The commodity used in this experiment was flowers with each individual initially asked to specify a *WTP* value to have flowers delivered to their home and a *WTA* value to perform this service for N other households. This approach was based upon previous work eliciting *WTP* and *WTA* values through contingent valuation methods (see Brookshire, et. al., 1986, for an overview). The low-bid auction procedure was then utilized to elicit a comparative set of *WTP* and *WTA* values. Thirty respondents were interviewed for each of three situations ($N = 5$, 10, and 50) where costs of delivering flowers were measured in terms of amount of time required to perform the task. These times were estimated beforehand as $M = 40$ minutes for $N = 5$, $M = 50$ minutes for $N = 10$, and $M = 90$ minutes for $N = 50$.

The results reveal that roughly 25 percent of all the respondents had values of X_i which were within 10 percent of the max-min estimates based on their initial *WTA* and *WTP* using (1). Although only one-third of the participants said they were willing to participate in the low-bid auction if played for real, over 80 percent felt the process to be a "reasonable one." An analysis of demographic variables of individuals conform to economic predictions. Individuals with higher opportunity costs of time reflected by higher wage rates required more compensation. The required compensation is increasing but diminishing with respect to N, because of increasing returns to scale based on time to deliver flowers.

III. Political Realities of the Process

The low-bid auction procedure must be viewed as one of a set of policy tools for dealing with the noxious facility siting process. It does offer the possibility of clarifying the relative costs and benefits of alternative locations and the appropriate monitoring and control procedures for implementing changes from the current system. By introducing competitive bidding, each community can determine whether it is worth entering the consortium by examining their default option in relation to their potential outcomes from specific *WTA* values.

Certain standards may have to be imposed by regional or state governmental agencies so that residents in all potential sites are convinced that they are sufficiently protected against adverse environmental effects such as pollution and noise. If the facility requires transport of materials such as when locating a trash plant, then additional regulations will have to be imposed on all delivery routes. In addition, some type of compensation arrangement may be needed to pacify those people located along the transport routes or those communities abutting on the site. The

actual amount could be prespecified in advance as a function of the lowest bid or it could be a fixed amount determined by the regional authority. Part of these costs could be covered by the budget surplus induced by the low bid auctions.

In the above example we assumed that all communities could host the facility. If certain parts of the region were not eligible as a site due to special features (for example, topographical or groundwater conditions), then they would be required to pay a tax based upon some prespecified criterion (lowest bid; average bid). If it was impossible to determine eligibility before undertaking specific geological studies, then each potential site could enter the bidding and would have to pay $1/N$ of its *WTA* unless it were the lowest bid and declared ineligible for other reasons. Under this arrangement, only the lowest bid site would be subject to detailed environmental and geological studies.

The low-bid auction supplements any arrangements between other interested parties in the siting debate. For example, the developer will have to specify the costs of building a facility at any one of the potential locations. Those sites which require a lower expense will have a comparative advantage in the bidding process since low-bid is defined as *WTA* plus cost of building the facility.

Finally there will undoubtedly be disagreements within the community as to *WTA* for hosting the facility. Public participation by all residents needs to be a part of the process to clarify the issues, but the low-bid procedure can only work if there is some collective choice mechanism such as a referendum by the citizens or the town government for finalizing a single bid.

IV. Extensions and Directions for Future Research

The above procedure needs to be modified if there are possible coalitions which can be formed so that one or more facilities can be located in an area. It is not obvious what type of iterative procedure would be applicable for this problem even if each participant utilized a max-min solution. For the case where a community must decide whether or not to participate in the auction, then the following sequential procedure could be utilized if there were only a single facility: have each of the N communities decide whether they want to join the cooperative or be excluded with a predetermined default option. If $N-j$ participants decide to join, then the *WTP* is increased from $1/N$ to $1/(N-j)$ of *WTA*. This may induce one or more participants to withdraw from the consortium. When the number of participants is the same for two iterations, then the low-bid auction is administered in the same manner as described in Section I.

An assumption fundamental to the above analysis is that only aggregate *WTA* or *WTP* of a community matters in bidding and selection. When the preferences of each community member are quasi linear in wealth and other goods, it is well known (see, for example, T. Groves, 1979) that aggregate *WTA*/*WTP* is a reasonable representation of overall community preferences provided that monetary transfers are possible within each community. Future research is needed to determine whether aggregate *WTA*/*WTP* is an appropriate basis for bidding mechanisms under more general conditions on individual preferences.

The low-bid auction can also be used under conditions of risk. This is important since some noxious facilities (for example, hazardous waste incinerators) may entail real or perceived risks if located in one's backyard. Under risk, one need only interpret V_i as the certainty equivalent (or option value) of the expected disutility of hosting the facility. In applications, one would expect insurance and other risk-spreading instruments to play a role as well, with V_i representing the net residual expected disutility from siting at location i.

In applying this procedure to siting problems involving risk, however, one must be mindful of recent findings in the literature on the psychology of choice. For example, recent controlled laboratory experiments and field studies suggest that the way one presents information on risk will influence the *WTA* and *WTP* (V. K. Smith and W. Desvousges, 1986). These findings are con-

sistent with framing and context effects related to risk as explored by D. Kahneman and A. Tversky (1981). Whether market mechanisms will correct for these difference is still an open question. Preliminary experimental results in asset trading suggest that in repetitive market-like environments subjects behave more in accord with economic rationality principles than experimental studies of individual behavior might suggest (P. Knez et al., 1985).

It should be clear that we have only begun to scratch the surface as to how one can locate noxious facilities where there are positive externalities to many and costs to a few. The problem is of sufficient practical importance as well as theoretical interest to challenge both researchers and policy analysts to explore other ways of utilizing compensation to improve the decision process as well as final outcomes over the relevant default options.

REFERENCES

Brookshire, D. S., Coursey, D. L. and Schulze, W. D., "Experiments in the Solicitation of Private and Public Values: An Overview," in L. Green and J. Kagel, eds., *Advances in Behavioral Economics*, forthcoming, 1986.

______, ______, **and Kunreuther, H.,** "Compensation Schemes in the Presence of Negative Externalities: A Field Experiment," 1985.

Clarke, F., "Multipart Pricing of Public Goods," *Public Choice*, Fall 1971, *11*, 17–33.

Cummings, R. D., Brookshire, D. S. and Schulze, W. D., *Valuing Public Goods*, Totowa: Rowman & Allanheld, 1985.

Groves, T., "Efficient Collective Choice with Compensation," *Review of Economic Studies*, April 1979, 46:227–41.

Kahneman, D. and Tversky, A., "The Framing of Decisions and the Psychology of Choice," *Science*, 1981, *211*, 453–58.

Knez, P., Smith, L. V. and Williams, A. W., "Individual Rationality, Market Rationality and Value Estimation," *American Economic Review Proceedings*, May 1985, *75*, 397–401.

Kunreuther, H. et al., "The Role of Compensation for Siting Noxious Facilities: Theory and Experimental Design," Risk and Decision Process Center Working Paper No. 85-04-03, Wharton School, University of Pennsylvania, 1985.

Morell, D. and Magorian, C., *Siting Hazardous Waste Facilities: Local Opposition and the Myth of Preemption*, Cambridge: Ballinger, 1982.

Smith, V. K. and Desvousges, W., "Asymmetries in the Valuation of Risk and the Siting of Hazardous Waste Facilities" *American Economic Review Proceedings*, May 1986, *76*, 291–94.

Tideman, T. N. and Tullock, G., "A New and Superior Process for Making Social Choices," *Journal of Political Economy*, December 1976, *84*, 1145–59.

The Regional Transformation of the American Economy

By BENJAMIN CHINITZ*

Let me begin by defining my regions. The 48 states are grouped for most statistical purposes into nine regions—these nine regions are further aggregated to form four large regions: the Northeast (the New England states and the Middle Atlantic states); the North Central (which in common parlance is called the Midwest and includes the industrial East North Central states and the agricultural West North Central states—sometimes referred to as the Plains states); the South (the south Atlantic, East South Central, and West South Central, a vast region stretching from the Atlantic Ocean in the East more than halfway across the continent); and finally the West (the Mountain states as well as the Pacific Coastal states, also a vast region, comprising 40 percent of the land area of the continent).

In my review of regional change, I will deal with three comprehensive measures of growth: population, employment, income. The years, 1900, 1950, and 1980 are particularly interesting for the wonderful reason that the total population of the United States was close to 75 million in 1900, increased by 75 million between 1900 and 1950, and increased again by 75 million between 1950 and 1980. The South and the West combined accounted for 38 percent of the population in 1900 but took 51 percent of the growth between 1900 and 1950, increasing their share of the U.S. population to 44 percent in 1950. In the next interval; 1950–80, they took 67 percent of the growth, increasing their share of the U.S. population to 52 percent. The West, which had only 5 percent of the population in 1900, took 25 percent of the growth in the twentieth century.

Thus, the combined impact of demographics and economics in the twentieth century has been to undermine the dominance of the older regions to the point where they account for less than half of the "action," both economically and politically. Four of the last five presidents brought to the White House a background in southern or western politics.

Gross employment figures by region add nothing to the overall perception of regional shifts conveyed by population figures, because the ratio of employment to population does not vary significantly by region. But personal income is another story. Here, let me focus on the slow growing regions of the North. In 1900, these regions accounted for 76 percent of personal income in the United States. That figure declined to 49 percent in 1980! Their share of the nation's income declined faster than their share of the nation's population. On a per capita basis they were much richer than the nation as a whole in 1900, but their position converged down towards the national average.

Two questions should be addressed at this point: *how* and *why*? The first relates to process; the second to cause. Let me offer you three distinct but complementary scenarios. I call them the Western, the Southern, and the Northern scenarios. To anticipate a bit: the Western scenario is one of sheer *growth in numbers*; the Southern scenario is one of *industrialization*; the Northern scenario is one of *adjustment to adversity*.

The Western Scenario: There are three ways a region's population can grow faster than the average for the nation: it could have a

†*Discussants*: Joseph Turek, State University of New York-Albany; Bennett Harrison, Massachusetts Institute of Technology; Gerald A. Carlino, Federal Reserve Bank of Philadelphia.

*Dean, College of Management Science, University of Lowell, Lowell, MA 01854.

higher rate of natural increase, it could be a favored destination for immigration from abroad, and finally, it could be favored as a destination for migrants from other regions. All three factors have contributed to the rapid growth of the West in the twentieth century, but the single most important factor has been interregional migration.

The West, because of its rich natural resources, vast open spaces, scenic beauty, and (for the most part) mild climate, has been a magnet to movers from the beginning of our history. It was always richer in per capita income terms and full of opportunity. But, it was not until 1900 that it commanded 5 percent of the nation's population as compared to almost 20 percent today.

The twentieth century has been good to the West in a variety of ways. I would emphasize transportation and communication. In the nineteenth century, the deliberate public policy to subsidize the construction of East-West rail lines encouraged westward migration to exploit its rich agricultural resources. But moving to the West was as traumatic a migration as sailing from English shores to the colonies was in the seventeenth and eighteenth centuries. Now, thanks to jet air transportation, automobiles, and "dual carriageways," long-distance telephone, television, and teleconferencing, a move to the West, both from a personal and a business point of view, does not sever contacts or relationships with the East. One can have one's cake and eat it too. The incredible developments in transportation and communications technology have unleashed the latent potential of the West for growth and development, a potential which has been enlarged by related other developments in the twentieth century, such as increased expenditures for travel and recreation, the emergence of defense expenditure as a significant proportion of *GNP*, and the creation of a class of retired elderly with money to "burn" in search of a mild climate and amenities. In the jargon of economics, the story of Western growth is one of "demand-pull."

The Southern Scenario: The more fascinating aspect of growth in the South is what I call the Southern scenario, which is fundamentally a "cost-pull" scenario. The South, as I define it, was "larger" than the Northeast in 1900, accounting at that time for 32 percent of the U.S. population as compared to the 28 percent for the Northeast. But while the Northeast boasted 41 percent of the country's personal income, the South was so poor relative to the North that its 32 percent of the population only enjoyed 15 percent of the nation's income! The South barely increased its 30 percent share of the U.S. population in the ensuing 80 years, but it brought its share of the nation's income up to the same level, doubling its 1900 share of 15 to 30 percent in 1980.

What was responsible for the vast improvement of income levels in the South? The single most important factor was the growth of manufacturing employment. Throughout the twentieth century, while the southern share of the U.S. population has been in the range of 30–33 percent, the southern share of U.S. manufacturing employment has persistently increased. In 1900, it stood at 15 percent—roughly half of the population share. In 1980 it was 30 percent, very close to its population share. Does this comparison strike a familiar chord? It is almost identical to the trend in population and income shares.

I call the Southern scenario cost-pull because the dramatic growth of manufacturing employment in the South, particularly in recent decades, has been achieved mainly at the expense of the North which used to be referred to as the nation's manufacturing belt. The South competed successfully with the North because of lower labor costs and generally lower operating costs. Lower labor costs were achieved not through higher productivity, but through lower wages. Labor was available at lower wages because labor was being released from agriculture, a sector in which sharp increases in productivity were being achieved through technology and capital investment. Between 1950 and 1980, employment in agriculture in the South fell from 3.2 million to 900 thousand. Throughout this period, manufacturing wage rates in the South have been in a range of 10–20 percent behind the North. The South has also been much less unionized than the North.

In the Western scenario I emphasized the pull of natural assets available in plentiful quantities. In the Southern scenario, I em-

phasize the pull of human assets available in plentiful quantities. What both scenarios share in common are the changes in technology that facilitated the exploitation of these assets. As transportation and communication became swifter and cheaper and more ubiquitously available, the opportunities for locating production facilities so as to tap into available supplies of labor at favorable rates were enhanced. There were more locations to choose from and a firm could locate those operations which were susceptible to cost savings in the South while maintaining other functions in the North closer to the centers of commerce and finance.

Wage rates aside, producers found many southern locations attractive in terms of other costs as well. There is ample evidence to suggest that tax costs—other than federal which tend to be uniform nationwide—were lower because state and local governments in the South extracted less revenue from their relatively poor constituencies and lagged behind northern governments in the financing of quality public services to all classes and welfare support and services for the poor. Furthermore, many of these governments offered subsidies to prospective employers in the form of lower taxes, cheap capital, physical facilities, and manpower training.

The absence of high-quality public services was undoubtedly a constraint on some forms of industrial development and qualifies the view that the South was a paradise to all entrepreneurs in search of low-cost locations. But on balance, the public sector added to rather than subtracted from the attractiveness founded on lower labor costs.

The Northern Scenario: I am struck by the North's resilience in the face of strong trends favoring the South and West. One might have expected the North to "unravel" in proportion to its losses of manufacturing employment, initially in terms *of share* and more recently in absolute numbers. After all, conventional theory tells us that the growth of a region is proportional to the growth of its economic base. Yet between 1950 and 1980, when manufacturing employment increased only 3 percent in the Northeast, total employment increased 39 percent, a ratio of 13 to 1, a ratio far greater than had been experienced in any region throughout the history of the country. In the South and the West, between 1950 and 1980, the growth of total employment actually lagged behind the growth of manufacturing employment.

The "salvation" of the North lay in the structural transformation of the national economy. In yielding manufacturing employment to the other regions, the North was, in effect, withdrawing from a sector of the economy which was shrinking as a share of the total economy and making room for growth in other sectors which were on the ascendancy in terms of their share of the total economy.

National growth in these sectors has "sustained" the North for two reasons. First, even a shrinking share of a growing pie can yield a bigger piece of pie. These sectors grow faster in the South and West than they do in the North, but they do grow in the North even in the face of declines in manufacturing employment in the North. Second, and most important, while employment in these sectors tends to be fairly evenly distributed among the regions in proportion to population and income, the North, on balance, enjoys a modest export surplus in these sectors so that its share of employment in these sectors, while falling, tends to stay above the region's share of population and income. The margin is slight as compared to the margin the region enjoyed in manufacturing in its heyday when its share of the nation's total manufacturing employment greatly exceeded its share of population. But a small margin on a broad base translates into a lot of jobs.

Let us consider two sectors which account for 34 percent of national employment and embrace a great variety of activities from barber shops to business consultants. I do not have the detailed data to support these claims, but it is reasonable to assume that cities such as New York, Chicago, and Boston export financial services to other regions; that Boston, New York, Philadelphia export health care and university education and research; that New York exports television and advertising and opera; that New England and the Great Lakes region export recreation. Of course, northerners also "im-

port" services from other regions, but the gross data do suggest the export surplus which we attribute to the North.

The point extends beyond these two sectors to the aggregate of nonmanufacturing employment. Consider the "bloody" period, 1960–75, when the Northeast region lost 781,000 manufacturing jobs while the South and the West gained 2,000,000 jobs. During that same interval, the Northeast *gained* 2,800,000 nonmanufacturing jobs, a lot less than the South and West did, but a lot more than you might have expected considering that Northeast manufacturing was virtually in a "free fall."

I describe this pattern of adjustment in positive terms, but I would not want to convey a sense of total complacency about the negative developments in the manufacturing sector *nor* do I want to leave the impression that all of the adjustment has occurred outside the manufacturing sector. On the first point, the northern regions in sequence—first, New England, then, Middle Atlantic, then, East North Central (Midwest)—have paid a heavy price for the erosion of their traditional economic base. Unemployment rates, particularly during cyclical downturns, were generally much above national averages. The great Northern cities which had done so well in the late nineteenth and early twentieth century in integrating waves of immigrants into the economy could not quite do justice to the multitude of blacks and Hispanics who joined their jurisdictions in the years after World War II.

Some of their problems were associated with suburbanization but the sluggish growth and ultimate decline of manufacturing employment in the region as a whole further drained cities of employment opportunities appropriate to the successful development of these minorities.

State and local governments in the region came under severe fiscal pressure as their revenues diminished and their costs kept rising. The agony was not restricted to the big cities and minority populations. All over the region, in places like Lowell, Lawrence, Youngstown, Johnstown, Scranton, as well as dozens of small towns and cities where the major source of employment had been manufacturing, plants were cut back or closed down, and there were pockets of high unemployment and poverty.

On the second point, there is adjustment within the manufacturing sector in the North from declining industries to growing industries, for example, from textile to computers in New England.

In the past ten years, New England has turned around dramatically. Between 1975 and 1981, manufacturing employment increased 10.7 percent in the United States and 17.0 percent in New England! (In the most recent recession, manufacturing employment in New England declined at a slower rate than any region of the country.)

In the same interval manufacturing employment in the Middle Atlantic region declined by 1.2 percent and only increased 1.5 percent in the East North Central region. In the 1980's so far, the most serious losses have occurred in the Midwest (East North Central) while the hemorrhage in the Mid-Atlantic region seems to have been arrested.

As I view these patterns, there is a suggestive sequence of growth, decline, stability and a resumption of growth. New England has completed the whole cycle. It is tempting to speculate that its sister regions in the North will achieve the same kind of recovery.

A Multiregional Model Forecast for the United States Through 1995

By BENJAMIN H. STEVENS AND GEORGE I. TREYZ*

To explain past, and predict future, regional economic performance in the United States over the 1967–95 period, we have constructed an eight-region model with 53 sectors per region. Output in each of 49 private nonfarm industries is divided into goods and services for intermediate use and final consumption within a region, and exports from the region, both interregional and international. Changes in exports for each of the 2-digit SIC industries are explained by changes in relative regional production costs, growth in interregional and international markets, and regional 3-digit industrial mix. Changes in region-serving output are generated endogenously by changes in requirements for intermediate goods and services, via input-output relationships; by changes in household, investment, government, and other final demands; and by changes in the extent of regional self-sufficiency. Changes in relative labor intensity and unexplained trend complete the equation for predicting employment in each sector.

Changes in wage rates, by sector, are explained by changes in labor demand by occupation, recognizing differences in wage responsiveness among skill groups, and by changes in overall labor supply due to changing population and participation rates. The occupational wage changes feed back to the wage bill (and to relative costs) of each sector via its occupational proportions; changes in the total wage bill affect incomes and consumption demands. Population changes are explained, in part, by employment changes.

The model for each region uses estimated parameters for responsiveness of business location to production costs, wages to labor market conditions, population to employment, and self-sufficiency to shifts in markets and supply sources. These parameters are estimated using industry specific and micro-data for all states. Major sources used to calibrate the model and to drive the forecast include the Bureau of Economic Analysis, for two-digit employment, wage and income data; the Bureau of Labor Statistics for 3-digit data and forecasts, Valerie Parsonick (1985); County Business Patterns for 4-digit data, plus the Census of Transportation, and a variety of other public sources. The methodology employed utilizes elements from computable general equilibrium, input-output, and macroeconometric modeling. Full details of the model structure and parameter estimation, and further references, will be found in our paper with Ann Friedlaender (1980) and our more recent article (1985).

In a short paper, it is possible to focus only on the explanation of differential changes in total regional employment shares (see Figure 1) using selected major causal factors aggregated over sectors. The major factors to be discussed here are: Relative Production Costs (Figure 2); Industrial Mix (Figure 3); Regional Self-Sufficiency (Figure 4); and Unexplained Changes in Employment Shares (Figure 5). Figure 1 shows the log of regional employment share minus the log of this value in 1967, thereby emphasizing deviations from constant rates of change. In all figures, FW = Far West, GL = Great Lakes, ME = Mideast, NE = New England, PL = Plains, RM = Rocky Mountains, SE = Southeast, and SW = Southwest.

Relative production costs in each region are calculated by forming a weighted average of relative costs in each of the 49 private nonfarm sectors. Total relative costs in each

*President, Regional Science Research Institute, Box 3735, Peace Dale, RI 02883; and Professor of Economics, University of Massachusetts, and President, Regional Economic Models, Inc., 306 Lincoln Avenue, Amherst, MA 01002, respectively.

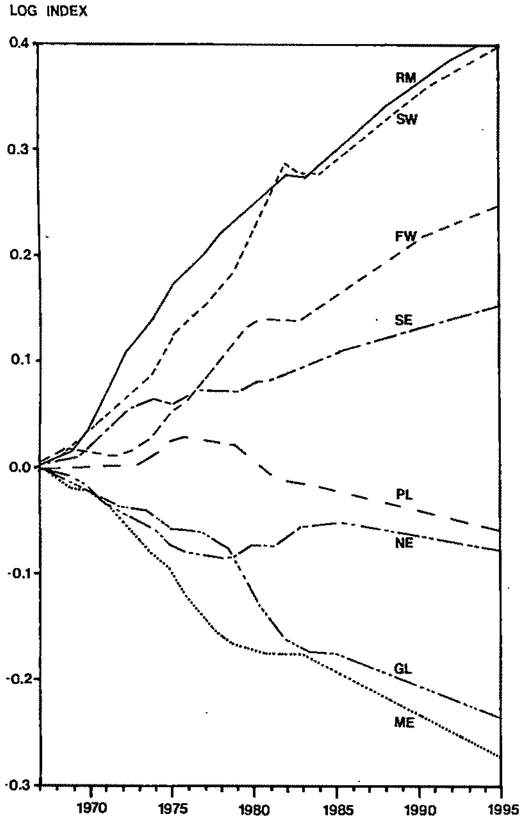


FIGURE 1. LOG INDEX OF REGIONAL EMPLOYMENT SHARE

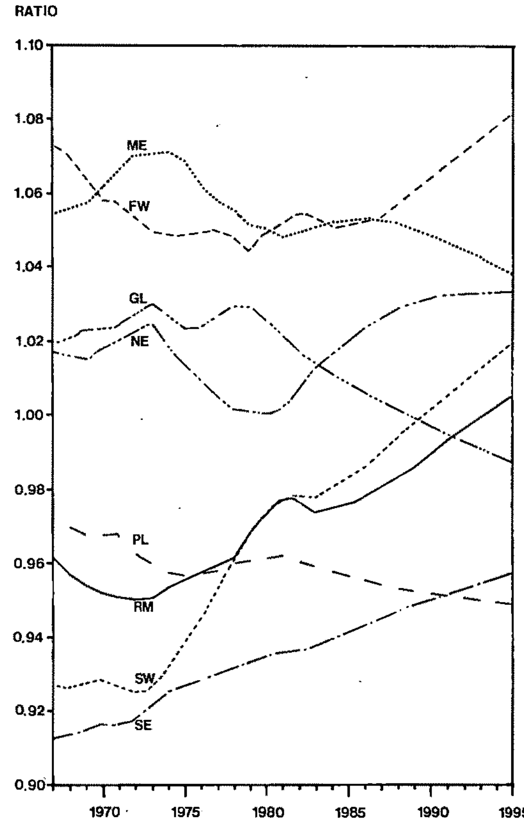


FIGURE 2. RELATIVE REGIONAL PRODUCTION COSTS

sector are calculated using labor costs, capital costs based on differential profit and property taxes, energy costs based on the regional prices and sector-specific mixes of three major types of fuel, and material and service input costs based on the region's relative costs in the sectors producing those inputs. Substitution among the aggregate KLEM factors is assumed to occur with unitary elasticity for all inputs except materials.

Figure 2 indicates that as might be expected, there is a tendency towards interregional cost convergence. Furthermore, comparison between costs and the changes in employment shares shows that three (RM, SW, SE) out of the four regions that have been growing relatively had below average costs in 1967. In addition, the two regions showing the greatest decline (GL, ME) are high-cost regions and the decline in New England costs in the 1970's preceded a revival of its economy in the 1980's.

However, there are some exceptions to the tendency toward cost convergence and to the expected inverse relationship between costs and growth. The Plains have lower than average costs which have shown no tendency to increase or to stimulate relative economic growth. The Far West has above average costs and growth in employment share in both the historical and forecast period. Thus, while relative costs may be an important influence on growth, they may be overwhelmed by other factors.

In Figure 3 we present an aggregate index that shows the effects of differential regional industrial mix on total employment share. The index shows the relative change in regional employment that would be forecast if the regions share of U.S. private nonfarm employment were held at 1982 levels.

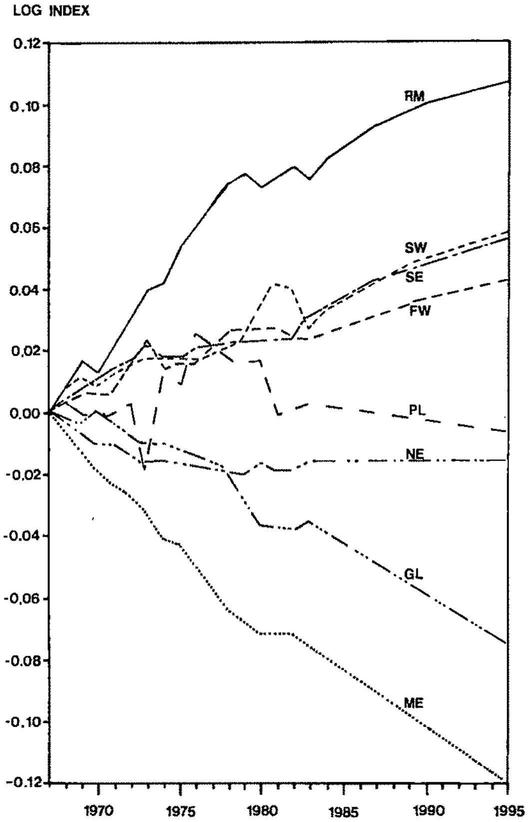


FIGURE 3. LOG INDEX OF RELATIVE INDUSTRIAL MIX EFFECTS

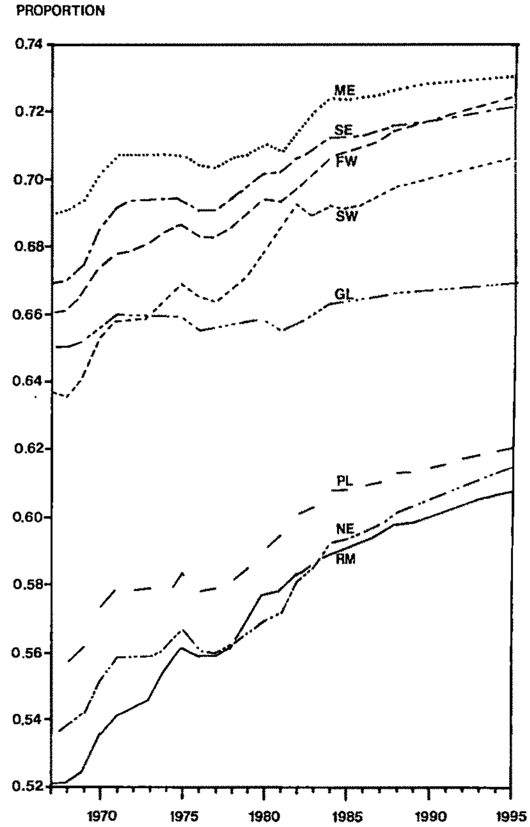


FIGURE 4. REGIONAL SELF-SUPPLY PROPORTION

The industrial mix is most favorable for three (RM, SW, FW) out of the four regions that are growing relatively but it is a negative factor for the fourth region showing relative growth (SE). The direct link between the deviations from trend of the industrial mix indices and the deviations from trend of regional employment can be seen by comparing Figures 1 and 4. This evidence supports the view that industrial mix plays an important role in both cyclical fluctuations and in longer-run trends.

A region's self-sufficiency is measured by a weighted average of its sector-specific regional purchase coefficients (*RPC*s); an *RPC* is defined as the proportion of the total use of a good or service in a region that is supplied by the region to itself. The weighted average *RPC*s can increase either by a shift in the industrial mix toward sectors with relatively high *RPC*s or by growth in individual *RPC*s themselves.

Figure 4 shows a tendency for all regions to become more self-sufficient over the period, with smaller regions starting lower and growing more rapidly in self-sufficiency. The most important factor leading to growing self-sufficiency is a general shift toward sectors with high *RPC*s, such as services. The other factor is growing regional diversification due, in part, to the increasing market orientation of manufacturing industries. This leads to increasing output to supply local markets through import substitution in industries that were previously underrepresented in the region. It is interesting to note that slow growth in the *RPC* measure in the Mideast and Great Lakes has apparently

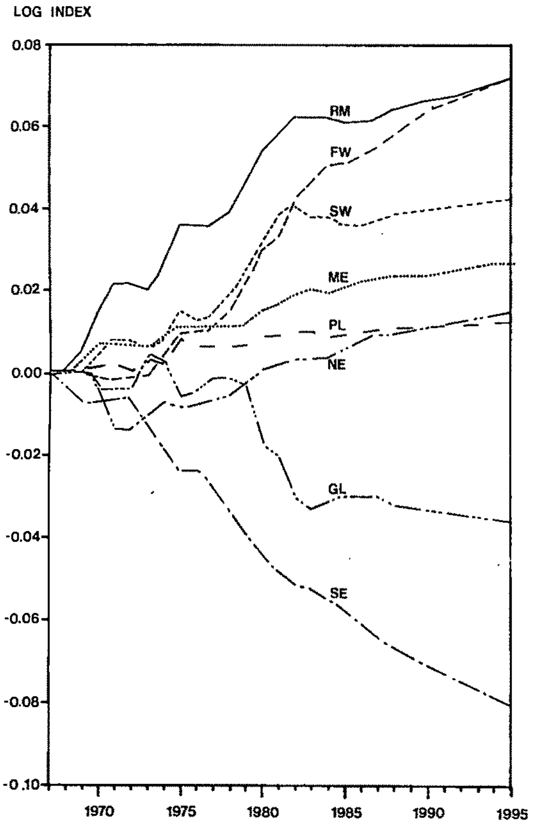


FIGURE 5. LOG INDEX OF UNEXPLAINED EMPLOYMENT SHARE

been contributing to the relative decline of these areas.

The unexplained change in employment shares (Figure 5) show remarkably stable trends, suggesting that the model will be successful in forecasting if these smooth trends are incorporated in the forecast. The most obvious exception to trend stability is the Plains region, whose erratic behavior in the 1970's appears to be associated with a period of rapid change in agriculture. Also a comparison of Figures 5 and 1 shows that relative employment changes that cannot be explained by the model are, in most cases, a small percentage of, and in the same direction, as the relative employment changes themselves.

The long-term stability of most of the trends, with unexplained growth in the South and West and unexplained decline in most of the North, supports the argument that the shift from the "Frostbelt" to the "Sunbelt" can only be partially explained by objective and measurable economic factors. Climate, of course, is often suggested as an additional cause. Sunbelt areas may attract migrants, such as retirees, with assets that allow them to consume at levels greater than the model predicts, based on their disposable income.

An additional, or alternative, possibility is that there is a growth psychology in growing regions. Even if the movement of employment, population and income were mutually determined, such movement might spur early investment in anticipation of further market growth and thus faster growth in self-sufficiency than our model predicts. This question deserves further study, especially since the regional shift toward the South and West are not forecast to abate significantly in the next decade. And these shifts will apparently continue despite the mitigating factors of declining costs and increasing self-sufficiency in at least parts of the Industrial North.

REFERENCES

Parsonick, Valerie A., "A Second Look at Industry Output and Employment Trends through 1995," *Monthly Labor Review*, No. 11, 1985, *108*, 26–41.

Treyz, George I. and Stevens, Benjamin H., "The TFS Regional Modeling Methodology," *Regional Studies*, 1985, *19*, 547–62.

______, **Friedlaender, A. F. and Stevens, B. H.**, "The Employment Sector of a Regional Economic Policy Simulation Model," *Review of Economics and Statistics*, February 1980, *63*, 63–73.

Analysis and Policy Implications of Regional Decline

By CHARLES L. LEVEN*

Urban economics is not a growth industry. Disinvestment seems evident in the shrinking of relevant flows (funded projects, students, new Ph.D.s) and stocks (course offerings, textbooks available). One view is that urban economics has failed as evidenced by the continued presence of poverty and blight despite decades of commitment to their elimination (R. Pfister, 1985). On the other hand, if one distinguishes between problems *in* cities, such as unemployment and crime, and problems *of* cities, like efficient land use and transport (my 1968 paper), it could be argued that cities have done quite well (A. Downs, 1981), even though people in them may not (N. Glazer, 1975).

However, whether one sees the urban condition as a "half-empty glass" or "half-full," it seems that macro theories of metro change have come up sadly lacking in explanatory power, especially since 1970. This is not because we have no theory of urban change: we do; it is called economic base theory. More complex formulations involve multisectoral regional accounts (see my 1961 article), interregional industry accounts (L. Moses, 1955), and or industrial complex analysis (W. Isard, 1975), but aside from niceties of differential impacts of or on particular sectors or nonlinearities, the essential mechanism in substance is equivalent to a Keynesian foreign-trade multiplier (C. M. Tiebout, 1962). It is captured even more simply in the concept of the nonbasic-basic ratio and the corresponding employment multiplier:

$$(1) \qquad [1+E_{NB}/E_B]\cdot dE_B,$$

where E_{NB}, E_B, and E are nonbasic, basic, and total employment, respectively.

Estimates of multipliers are sensitive to the specification of nonbasic vs. basic activities and techniques for estimating this split can produce different results. For example, it makes a difference how we define sales within the area to visiting nonresidents, how we account for sales between businesses in the area, and whether construction is treated as endogenous or exogenous. It also matters whether we determine the nonbasic-basic breakdown of a particular sector by survey, by use of localization coefficients, or by functional assignment, and aggregation bias can follow from the choice of sectoral specifications at which analysis is carried out.

Notwithstanding these difficulties, it has been widely believed that at least crude results could be obtained by defining manufacturing and possibly wholesale employment as basic and other employment, perhaps net of construction, as nonbasic (H. W. Richardson, 1979). This results in the calculation of the multiplier as

$$(2) \quad 1+[(E-C)-(M+W)]/(M+W),$$

where M, W, and C are manufacturing, wholesale, and construction employment, notwithstanding that some manufacturing and wholesale trade are for local markets and that services, retail trade, and utilities can be exported to some extent. In fact, this crude formulation is expected to work not just as a statistical convenience, but out of a theoretic argument that it should. Population agglomerations, traditional location theory tells us, are a consequence of scale and agglomeration economies in the production of those goods having nonzero bulk-to-value ratios (Isard, 1954). Services are defined as nontransportables not subject to scale economies in production and are produced at goods production locations in some more or less fixed ratio giving a multiplier definition of

$$(3) \qquad 1+(S/G),$$

where S denotes services and G denotes goods.

*Professor of Economics, Washington University, St. Louis, MO 63130.

If location theory is right and ignoring such technicalities as possible exports in retail trade to a commuting hinterland and sales of retail goods and services to visitors then (2) is pretty much a simple operational approximation of (3). However, since some components of services are partly transportable and to some extent subject to agglomeration, as opposed to scale economies, we should expect the value of (2) to increase with population of an urban area but remain fairly stable over time for a given area, except for scale effects.

So formulated, economic base theory largely fails to explain recent changes in urban development (see Table 1). In fact it does not even get the sign right. In all of the first six urban areas, a falling $(M+W)$ is associated with a rising $[(T-C)-(M+W)]$; in the remaining four areas it is falling. Defining the basic sector as manufacturing only and/or defining the nonbasic sector as including rather than excluding construction yields similar results.

There is a competing theoretical explanation, namely that total employment is a simple transform of exogenously determined population scale, with nonbasic employment determined as a function of local supply conditions for nonlabor inputs and the local wage rate, both of which are region specific. Basic employment, then, is determined as the residual of total less nonbasic employment, since assuming competitive markets there will always be an infinitely elastic demand for basic sector output at a fixed delivered price (G. Borts and J. Stein, 1964). In this competitive allocation theory, the critical explanatory factor explaining change is not export demand, but labor supply.

One application of this theory is to define areas in the Sunbelt as attractive and those in the Frostbelt as repellant; Sunbelt could be defined as the South and West Census Regions and Frostbelt as the Northeast and Midwest (V. Arnold, 1980). This definition admittedly is fairly crude in that there are some unattractive Sunbelt areas and some attractive Frostbelt ones. Nonetheless, if the theory is right, it should at least approximately explain differential change for these regions; it does not do so. Its failure is not due to the absence of net population shift from the Frostbelt to the Sunbelt, but rather that the rate of that shift after adjusting for urban hierarchy effects independent of region has not increased in recent decades and, if anything, has become less important in historic terms.

The *net* shift estimates in Table 2 were determined by a shift-share calculation where net shift is simply actual less expected growth for that decade and expected growth is what would have resulted if each cohort in each region grew at the national rate for that

TABLE 1—SMSA COMPARISONS, 1959–83

	Loss of[a] Mfg Jobs	Gain of Service Jobs[a]	Total Employment[b]	Transition Index[c]
Pittsburgh	−136	+118	+5.9	44
Cleveland	−74	+107	+21.1	32
Detroit	−128	+200	+23.1	34
Chicago	−147	+440	+32.9	32
Milwaukee	−42	+102	+37.5	38
St. Louis	−53	+155	+39.9	38
Minneapolis	+81	+170	+105.8	31
San Diego[d]	+39	+131	+184.2	42
Denver	+72	+140	+214.7	27
Houston[e]	115	+267	+246.3	31
U.S.[d]	+3365	+12786	+77.3	29

[a]Shown in thousands.
[b]Percent change.
[c]Percent decline in $1+[(T-C)-(M+W)]/(M+W)$.
[d]1982
[e]Excludes Liberty and Montgomery counties.

TABLE 2—SHIFT RATE BY REGION[a]

	1890–1900	1910–1920	1930–1940	1950–1960	1970–1980
NE	−37	−59	−40	−120	−91
SATL	−30	−60	−31	−93	−105
ENC	+1	+19	−22	+2	−78
WNC	−24	−33	−42	−15	−61
SATL	−18	+30	+56	+65	+69
ESC	[b]	−45	+19	−72	+7
WSC	−106	+63	+5	+22	+122
MTN	+237	+169	+61	+207	+264
PAC	+107	+148	+109	+148	+93

[a]Persons per 1,000.
[b]Less than 0.5 persons.

cohort. The figures in the table are expressed per thousand of initial population at the beginning of each decade. The shift-share analysis that produced these estimates was conducted at the state level with 11 urban hierarchy cohorts: central cities of large, medium, and small SMSAs; suburbs (SMSA less the central city) of large, medium, and small SMSAs; exurbs (counties adjoining SMSAs) of large, medium, and small SMSAs; small-town areas (any other county with a town of at least 30,000); and rural areas (remaining counties).

Large SMSAs are those with 50 percent, medium with 25 percent, and small with 25 percent of aggregate SMSA population cumulatively as of the beginning of each decade. The SMSAs prior to 1950 were as defined in an earlier study (S. Sheppard and myself, 1984). The no-trend-in-net-shift result is insensitive to the use of different SMSA-size class definitions, changing the small-town cutoff to 20,000 and combining small-town and rural counties.

These results are interesting for practice, but disastrous for the theory. That theory strongly suggests that net shift should have increased as a function of high-income elasticity of demand for amenity, historic fall in the price of air conditioning, and increased relative price of fuel in the Northeast. It has not. I conclude that neither economic base nor environmental amenity explanations of urban development are consistent with recent historical experience. They do not allow us to deduce meaningful scenarios for urban development in advanced industrial countries.

To derive a new scenario we need to have some hypotheses about what was wrong with traditional base or environmental attraction theories. Since they are not logically flawed, the problem must be in the underlying axioms concerning technical constraints or behavioral response. Three possible hypotheses seem likely candidates for explaining the apparent contradictions, (see my 1985 article).

One possibility is that demand for amenity actually determines location of population; location of population determines nonbasic employment; and population and nonbasic employment jointly determine basic employment. The aspect of amenity affecting locational behavior is not the fixed resource characteristics of a location, but the range and supply price of the nonbasic components themselves. While these may be affected by natural causes, it is mainly the locally determined institutions for making nonbasic commodities available, especially those subject to scale economies in consumption that are the important exogenous element. This scenario would be driven by a Borts-Stein mechanism, but "attractiveness" would be a function partly of yesterday's attractiveness and partly of public institutional response; we would look to history and politics rather than to climate and resources as crucial, and the core of the theory would refer to determinants of the consumption vector. Within this reformulation it is the "central placeness" of an area which is its most important amenity. If this scenario is a major part of what has happened to the urban process, we should expect a flattening of the size distribution of SMSAs (service specialization would be much less idiosyncratic than goods specialization) but substantial interregional neutrality (see my 1979 paper).

A second possibility is that demand for exports is still critical but that intangible services have replaced tangible goods as the larger component of basic activity due to two related changes in information technology. First, intangibles themselves may have become somewhat more subject to scale economies in production; even more they may have become much easier to transport. Here the scenario would derive from spatial agglomeration of intangibles; unfortunately we have no operational theory of the location of intangibles, since it would rely on estimates of interregional differentials in unit costs of inputs (mainly labor services), which decidedly are only very imperfectly reflected in wage rates. When this process might have started could be indicated by shifts in the distribution of service production among areas of different population size, though since obsolescence of both largest and smallest areas is likely changes only in a Gini coefficient would be inconclusive.

The third possibility is that what has really changed is neither demand for or the loca-

tional determinants of intangibles per se, but the automation of the production of tangible commodities. In particular, we may be witnessing an increase in information intensity (a form of capital deepening) of goods production with an expansion of services as intermediate inputs to goods production. This scenario would be driven by a secular fall in the multiplier for an area of fixed-population size and the theory required would have as its main argument elasticities of substitution of intangible for tangible inputs in the production of tangibles. Unfortunately we do not have interindustry data that permits this kind of analysis at a meaningful level of sectoral disaggregation. A crude index of multiplier shift for selected areas is shown in the last column of Table 1.

Most likely, all three kinds of structural change are involved. It is probably not just changing technology, since services as a proportion of *GNP* began rising well before the microchip revolution. On the other hand, recent population turnaround in small-town areas and out-migration from almost all large SMSAs suggest a post-computer-chip substitution of off-site service for on-site operative inputs in the production of tangibles.

Whatever combination of these three scenarios is at work, it seems clear that they have produced a very distinct improvement in the urban condition; property abandonment has ceased, gentrification if of modest proportion continues, talk of municipal bankruptcy has faded and central cities are experiencing rebirth; while investment tax credits may explain much of the recent boom in investment, they do not explain its resurgence in traditional downtown settings in older SMSAs.

Of course, everyone has not benefited from metro turnaround. It has been a process involving a return of jobs much more than a return of people, so that many of the underserved and isolated innercity poor may be as ill-served as ever. Also, job growth has involved a substantial shift from blue- to white-collar work, from larger to smaller production units, and from more to less organized regimes, including growth in off-site work, even in workers' homes. This has lowered the demand for the traditionally organized blue-collar workforce, many of whom have experienced loss of relative earnings, or even employment itself. We should not be surprised that they and their advocates resist and would even like to reverse the transition through impediments to business relocation or closure, especially for industrial plants (B. Bluestone, 1984). An even larger issue is the potential polarization of society as jobs for skilled professionals and unskilled service operatives replace jobs for moderately skilled craftsmen (R. Louv, 1985). As economists we can point out the conflicts between efficient resource transition and inequities to owners of particular specialized inputs (in this case, mainly skilled blue-collar workers) and we can demonstrate the welfare superiority of compensating the losers over inhibiting the transition. We also can understand why, if compensation does not take place, that some might strongly advocate process interference even if it were inefficient in the aggregate.

Indeed, it may be that recent success of the U.S. economy is due as much to successful transition of its cities to a post-industrial status as it is to bungling but expansionary monetarism of sound money and lots of it. The British economy, on the other hand, clings to the notion of recapturing industrial export markets either through trade restrictions (Labor) or lowered industrial wages (Thatcherites). It may just be that their only choice is between fairly high wages for a much smaller number of industrial workers, or high employment at very low wages. Urban economics, it seems, does have very much of a future, not necessarily in arbitrating between the philosophical advantages of market vs. planned solutions, or workers vs. rentier interests, but rather to carefully map out the possible choices and options for all groups and ways to advance the locus of choices for all.

In the United States, much of the significance of the efficiency-equity conflict may be beside the point to the extent that the reallocation already has taken place. Given that manufacturing employment already has declined to about 20 percent of the labor force, and given that a large share of those so classified really are information not materials handlers, most of the transition may already

have occurred (see Pittsburgh in Table 1). In the next century, historians may look back to the 1990 Census as the first one that looked at the tangibility of their inputs as perhaps a more significant characteristic of workers than the tangibility of their output; and it may be the last Census to regard the distinction between metropolitan and nonmetropolitan counties as important.

REFERENCES

Arnold, V., *Alternatives to Confrontation*, Lexington: Lexington Books, 1980.

Bluestone, B., "Is Deindustrialization a Myth? Capital Mobility vs. Absorptive Capacity in the U.S. Economy," *Annals of the American Academy of Political Science*, Spring, 1984.

Borts, G. and Stein, J., *Economic Growth in a Free Market*, New York: Columbia University Press, 1964.

Downs, A., *Neighborhoods and Urban Development*, Washington: The Brookings Institution 1981.

Glazer, N., "Planning and Policy Dimensions," in G. Sternlieb and J. W. Hughes, eds., *Post-Industrial America: Metropolitan Decline and Inter-Regional Job Shifts*, New Brunswick: Center for Urban Policy Research, Rutgers, 1975.

Isard, W., "Location Theory and Trade Theory: Short-Run Analysis," *Quarterly Journal of Economics*, May 1954, *68*, 305–20.

______, *Introduction to Regional Science*, Englewood Cliffs: Prentice-Hall, 1975.

Leven, C. L., "Regional Income and Product Accounts: Construction and Application," in W. Hochwald, ed., *Design of Regional Accounts*, Baltimore: Johns Hopkins Press, 1961.

______, "Towards a Theory of the City," in *Urban Development Models*, Highway Research Board, Washington: National Academy of Sciences, 1968.

______, "Economic Maturity and Metropolis' Evolving Physical Forms," in G. A. Tobin, ed., *The Changing Structure of the City: What Happened to the Urban Crisis?*, Beverly Hills: Sage Publication, 1979.

______, "Regional Development Analysis and Policy," *Journal of Regional Science*, November 1985.

Louv, R., *America II*, New York: Penguin Books, 1985.

Moses, L., "The Stability of Interregional Trading Patterns and Input-Output Analysis," *American Economic Review*, December 1955, *45*, 803–32.

Pfister, R., "U.S. Urban Policies: A History of Failure," 32nd North American Meetings of the Regional Science Association, Philadelphia, November 16, 1985.

Richardson, H. W., *Regional Economics*, Urbana: University of Illinois Press, 1979.

Sheppard, S. and Leven, C., "Changing Structure of the Metropolitan Dimension of U.S. Population: An Historical Perspective," IURS Working Paper RPS5, Washington University-St. Louis, June, 1984.

Tiebout, C. M., *The Community Economic Base Study*, New York: Committee for Economic Development, 1962.

Corporate Control, Insider Trading, and Rates of Return

By HAROLD DEMSETZ*

The attention focused on the role of corporate takeovers in disciplining management has led to the neglect of a more important method for joining management and owner interests—an ownership structure appropriately concentrated under normal circumstances to deliver an efficient level of monitoring. A study of ownership concentration in over 500 very large corporations by myself and Kenneth Lehn (1985), referred to hereafter as the D-L study, discovered that the 5 largest ownership interests controlled over 25 percent of shares in 1980, and that differences in concentration across firms in the sample respond to the benefits and costs of monitoring management. This reflects the existence of a monitoring mechanism more basic to and more continuously operating than the corporate takeover.

It was also observed from this sample that when individual and family ownership interests owned large fractions of a firm, these same interests failed to own significant amounts of shares in other firms. Important institutional owners did not specialize their portfolios in this fashion, nor did they own such large fractions of a firm's shares. Both differences may reflect the legal and fiduciary constraints facing institutional investors. It is doubtful if there would exist much incentive to monitor management in the absence of investment specialization of the sort exhibited by dominant individual and family owners. For this reason, I shall identify individual and family owners who own relatively large blocks of shares in any one firm as *controlling* shareholders of that firm.

This 511-firm sample tells us nothing about other investments that might be made by controlling shareholders, but I ask the reader to assume that the specialization of investments revealed in this sample correctly identifies a tendency by them to concentrate the uses to which they put their wealth. The tendency to specialize investment in a single firm burdens them with a cost not borne by minority shareholders—firm-specific risk. Another finding of the D-L study indicates that controlling shareholders bear even more firm-specific risk; ownership is more concentrated in firms exhibiting higher firm-specific risk. What benefits motivate investors to become controlling owners in the face of greater firm-specific risk?

Undoubtedly, many persons own a large fraction of shares in a firm because they have, or feel they have, a comparative advantage in exercising control, and that this advantage is worth utilizing to realize pecuniary and nonpecuniary returns. This must be the primary explanation for ownership concentration of the stable variety that is of concern here. Controlling owners also may be descendants of the firm's founding family, locked by capital gains tax avoidance into retaining a controlling position, but this incentive disappears when an estate is transferred at time of death. The central questions of this paper are whether insider trading offers a secondary compensation to controlling shareholders, and, if so, what are its characteristics and consequences?

The greater the percentage of shares owned, the greater is the power to obtain representation on the board of directors and to exercise influence over the management team. This power, and the continuing con-

†*Discussants*: John J. McConnell, Purdue University; David W. Mullins, Jr., Harvard University.

*Professor, Department of Economics, UCLA, Los Angeles, CA 90024. Ken Lehn and the SEC cooperated in data provision and analysis. Carol Simon made useful comments, and Susan Woodward's insights about the impact of insider trading on stock rates of return motivated the last part of this study.

tact with the firm's affairs that it implies, provides access to insider information. Given the requisite large holding of the firm's shares, controlling shareholders are much more interested in having "their" firm experience good news rather than bad, so insider trading augments comparative advantage in encouraging the establishment of effective and beneficial control of management. Competition of substitute controlling shareholders determines the price of acquiring control. Combined with the inability of controlling shareholders to alter their substantial stock holdings quickly without dramatically affecting share prices, this serves to limit the return they can obtain.

Studies have shown, and we have every reason to believe, that trading profits are greater for traders possessed of insider information. For example, Joseph Finnerty (1976) shows that after insiders sell a stock heavily, its price falls by an abnormal amount and, after they buy heavily, its price rises by an abnormal amount. These studies, however, shed little light on the ability of controlling shareholders to tap insider information. It is so plausible that they have superior access that a demonstration seems unnecessary, but indirect evidence can be given. I base this on the expectation that greater access to more profitable inside information should lead to more active trading.

The SEC data on insiders for 1980 has been used to calculate insider trading volume as a percentage of insider stock holdings for two sets of firms drawn from the D-L study. The first subset contains the 28 firms exhibiting the highest degree of ownership concentration.[1] The second contains the 28 firms exhibiting such low concentration of ownership that no one shareholder owned more than .02 percent of the shares. If accessibility to more profitable insider information is greater for shareholders owning relatively more shares, the first subsample should show greater involvement in trading by those classified as insiders by the SEC. In fact, there is 7 times more involvement in insider trading. Shares actually traded by persons classified as insiders, as a percentage of shares owned, is .64 percent for the first subset. This is to be compared to .09 percent for the second subset. Access to insider information does seem to be related to importance of shareholdings. Actual insider trading volume undoubtedly exceeds these reported values, but probably has no significant effect on their relative sizes.

A direct test of the importance of insider trading profit as an incentive for maintaining controlling ownership positions would examine the relationship between *profitability* of insider trading and ownership concentration. This would avoid the problem of underreporting of insider trading *volume* to the SEC. Unfortunately, comprehensive data by firm on profitability of insider trading are not now available.

Firm-specific risk is itself a plausible measure of the profit potential of insider trading. High systematic risk firms are those whose fortunes in greater degree are tied to factors that also influence the fortunes of other firms. High firm-specific risk firms are those whose fortunes tend to be tied to factors that do not influence many other firms. Information about common factors, such as growing tightness in capital markets, will be known in advance to many persons in many firms that stay in contact with capital markets. Profiting from this information is difficult because intensive competition to do so is faced from all who are well positioned to have the same information. In contrast to this, advanced knowledge about a successful closing in a new large contract is more likely to be restricted to persons in firms doing the contracting. Trading on the basis of such firm-specific information is likely to be less competitive and more profitable. It is information that impacts the fortunes of a specific

[1] I arbitrarily chose 60 firms to measure differences in intensity of involvement in insider trading and 160 firms to measure the correlation between insider trading and firm-specific risk. Missing data forced me to delete firms, and the publishing deadline of the *Proceedings* barred the replacement of these firms with others. Hence, I use 28 firms exhibiting high-ownership concentration and 28 exhibiting low-ownership concentration to measure differences in the ratio of insider trading to shares owned by insiders, and I use 159 firms to correlate firm-specific risk and insider trading as a fraction of total trading.

firm that provides the best opportunity to profit. Such information is most frequently encountered in those firms exhibiting high firm-specific risk. In fact, calculations on a 159 firm subset of the D-L sample shows a high correlation between the ratio of insider trading to total trading (measured for 1980) and firm-specific risk (measured over the period 1975–80). The correlation is .45 (with a t of 6.4).[2]

For this reason, the statistically significant positive relationship between ownership concentration and firm-specific risk found in the D-L study indirectly supports the proposition that the supply of controlling owners is positively related to the potential profitability of insider trading. The positive relationship between ownership concentration and firm-specific risk is interpreted in the D-L study as reflecting a greater demand for monitoring in firms subject to greater instability in the immediate environments in which they compete. Firm-specific risk is taken as an index of this instability. On this view, greater firm-specific risk implies a larger demand for controlling shareholders. This demand cannot bring forth more controlling shareholders without compensating them for bearing greater firm-specific risk. The larger part of this compensation presumably comes from the production of more valuable firms through the exercise of controlling ownership, a compensation that would be greater in those firms facing the most unstable conditions. But it will also be true that insider trading profits should be greater in these same firms. This constitutes a secondary source of compensation.

That insider information allows insiders to obtain a higher than average rate of return in the market for shares has been documented in other studies, but the interpretation given to such findings is misleading. Insiders bear special costs insofar as they are controlling shareholders. The cost of firm-specific risk is not recorded as an expense of the firm. It is borne personally by controlling shareholders. The higher rate of return *recorded* by insider trading in the market for shares, therefore, cannot be interpreted to mean that controlling shareholders receive a higher *personal* return than is enjoyed by outsiders.

Of course, controlling shareholders are not the only persons who come into possession of advance information. Others with access to such information would include key employees, bankers, and others who do business with the firm. Are they profiting from insider trading without improving the firm's operations? Some of them undoubtedly are in any given episode of new information. But economists should be suspicious of free lunches. The appearance of one often reflects our inability to observe the price easily. In this case, it probably reflects our ignorance of what is required of persons seeking regular access to insider information. Key employees bear the risk of specializing their human capital to the needs of the firm. Bankers and others who do important business with a firm not only receive information from the firm, but they also bring information to the firm. Insider information is a way of paying for services such as these, which, in their very nature, are difficult to reward through explicit contracts.

Trading with insider information, of course, creates opportunities to profit at the expense of other traders. These wealth transfers, to the extent they impact minority shareholders, may be viewed as part of the cost borne by minority shareholders to encourage more effective monitoring of the firm. Since this cost falls in greater degree on minority shareholders who trade more frequently, minority shareholders should hold stock for longer periods than they would in the absence of insider trading. Buy-and-hold strategies will not give perfect protection against insider trading because they impose illiquidity costs on investors. Insider trading profits thus give rise to problems of equity, not only between insiders and "outsiders," but also between minority shareholders facing different liquidity costs. Equity problems such as these are a primary source of opposition to insider trading. But, legislation seeking to reduce insider trading profit makes it more difficult to maintain controlling ownership interests so useful for monitoring purposes.

[2]See fn. 1.

The wealth transfers associated with insider trading may be smaller than is thought if outsiders can anticipate which stocks confront them with higher probabilities of informationally disadvantaged trading. It is not clear that such anticipations can be formed confidently. Forecasts of insider trading must be made for periods long enough to make buy-and-hold strategies meaningful. Controlling stockholders, key employees, and persons with important business relationships all have access to insider information, so it is not clear that a relationship exists between insider trading and concentration of ownership. Although the general tendency of a stock to be traded intensely by insiders gives a plausible guide to future insider trading, the expected financial impact of insider trading on outsiders may be too small per trade to warrant careful investigation. All in all, it may be better simply to accept the average amount of insider trading that characterizes a diversified portfolio. To the extent that outsiders can and do make such forecasts, they can limit their losses to insiders by reducing their demand for stocks most likely to confront them with informationally disadvantaged trades. Investors would discount these stocks, or they would require higher dividends from them, so as to equalize rates of return *realized* after losses to insiders. This implication of efficient markets requires *recorded* rates of return, measured before losses to insiders, to be higher for stocks more likely to present outsiders with informationally disadvantaged trades. Difference in recorded rates of return will not be arbitraged away by outsiders because they are necessary to equalize realized rates of return.

Based on a 159-firm subset of the D-L data, there is a significant correlation between the recorded market rates of return, R, measured over the period 1975–80, and the ratio of insider trading, v, to total trading, V, measured for 1980 (t-statistics are shown in parentheses):

$$R = \underset{(15.8)}{.016} + \underset{(3.8)}{.0005} v/V; \quad R^2 = .08; \ F = 14$$

This result supports the notion that investors discount stocks traded in intensely by insiders, but it may also reflect the happenstance that a disproportionate share of insider trading was done in anticipation of good news. On the other hand, for these same stocks, no significant correlation exists between this measure of the importance of insider trading and intercepts of the market model (estimated for 1975–80 in the D-L study). This suggests further analysis is required to isolate the effects of insider trading from riskiness of investment, a task likely to be difficult if both reflect the rate at which new information impacts stock price.

REFERENCES

Demsetz, Harold and Lehn, Kenneth, "The Structure of Corporate Ownership: Causes and Consequences," *Journal of Political Economy*, December 1985, *93*, 1155–77.

Finnerty, Joseph E., "Insiders and Market Efficiency," *Journal of Finance*, September 1976, *31*, 1141–48.

Mergers, Buyouts and Fakeouts

By MARK HIRSCHEY*

Economists have long recognized the advantages to managerial specialization, but have often expressed concern for the divergence in economic interests which may arise between stockholders and managers with little ownership interest. Adolf Berle and Gardiner Means close their famous book *The Modern Corporation and Private Property* with fears of "corporate plundering" (p. 355) by management; behavioral theories of the firm emphasize the role of "expense preference" behavior by managers (Oliver Williamson, 1963); and the modern theory of finance accords a prominent role to the study of agency problems (Michael Jensen and William Meckling, 1976). Of course, predating these concerns is Adam Smith's anticipation of the agency problem for joint-stock companies where

> ...[T]he directors of such companies, however, being the managers rather of other peoples money than of their own, it cannot well be expected, that they should watch over it with the same vigilance with which the partners in a private copartnery frequently watch over their own....negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company.
> [1776, p. 741]

Despite this long history of interest and work on the subject, research in accounting, finance, economics, and law seems more intrigued than ever with issues surrounding a potential divergence in owner vs. managerial interests. The *Journal of Financial Economics*, for example, recently featured a symposium issue titled "The Market for Corporate Control" which included studies on the wealth effects of takeover activities, antitrust, and the sources of merger gains, and management vs. stockholder interests (Jensen, 1983). In addition, *The Journal of Law and Economics* featured a special issue based upon the "Corporations and Private Property" conference held in commemoration of the fiftieth anniversary of the publication of *The Modern Corporation and Private Property* —a book described therein as one of the most influential publications of the twentieth century (Thomas Moore, 1983); and *The Journal of Accounting and Economics* has featured a symposium issue on "Management Compensation and the Managerial Labor Market" (Jensen and Jerold Zimmerman, 1985).

In addition to its prominent and timely place on the academic research agenda, the subject is of obvious current policy interest. The 1985 *Economic Report of the President*, for example, devotes substantial space to policy concerns surrounding the market for corporate control. In support of much academic research on the topic, a conclusion is reached that activities in the market for corporate control have generated net benefits for the economy and that further federal regulation would be premature, unnecessary and unwise. Still, this policy conclusion, like much recent research on the subject, remains controversial.

The purpose of this paper is not to broadly survey and document the methodology and findings of recent studies on the market for corporate control. That task is ably met elsewhere (see Jensen and Richard Ruback, 1983). Nor is my purpose to adopt that uniquely academic approach of taking a timely subject of great current interest and rendering it less so. Instead, the purpose of this paper is to cast mergers, buyouts, and what I refer to as fakeouts, as complementary mechanisms that act to complete the market for managerial talent, and therein comprise important elements in the market for corporate control. In so doing, my debt to Henry Manne (1965), Eugene Fama (1980), and Michael Jensen becomes obvious.

*University of Colorado at Denver, Graduate School of Business Administration, Denver, CO 80202.

I. The Managerial Labor Market

Fama has offered theory suggesting the separation of security ownership and control can be explained as an efficient form of economic organization within an agency theoretic perspective. In particular, individual employees in the firm, including its management specialists, face both the opportunities and discipline of external and internal labor markets. Whereas stockholders have an obvious ownership interest in firm performance, management has a perhaps less obvious compatible interest in that a firm's performance record constitutes the best available indicator of management's economic productivity. In Fama's words, "...the managers of a firm rent a substantial lump of wealth—their human capital—to the firm, and the rental rates for their human capital signaled by the managerial labor market are likely to depend on the success or failure of the firm" (1980, p. 292). Since firm performance is determined, at least in part, by the performance of the entire management "team," each manager has a stake in the performance of managers above and below them and will actively engage in two-way monitoring. Therefore, opportunity cost wages signaled by the external labor market and the two-way managerial monitoring of the internal labor market work together to discipline managerial performance and eliminate self-dealing or agency costs.

Fama's model of internal and external monitoring that stimulates managerial performance and the ongoing efficiency of the corporate form can be regarded, in the nomenclature of finance, as a strong-form theory of managerial labor market efficiency. In this view, the human capital and human capital rental rates of managers reflect all information concerning management performance, whether publicly available or not. As such, strong-form theory abstracts from the likelihood, or even possibility, of asymmetric information pertaining to managerial performance.

In operationalizing the notion of capital market efficiency, Fama (1970, 1976) defines three different levels or types of efficiency. Each of these is based on a different view of what constitutes relevant information in the capital asset pricing process. In weak-form efficiency, no excess returns can be earned based on trading rules derived from past price or return data. In semistrong-form efficiency, no excess returns can be earned based on trading rules shaped by publicly available information. Strong-form efficiency requires that no excess returns can be earned using any public or private information regarding firm performance. Despite troubling anomalies, empirical capital markets research provides support for both weak-form and semistrong-form versions of the capital market efficiency hypothesis, while rejecting the strong-form efficiency version (G. William Schwert, 1983).

The first main thesis of this analysis is that, by virtue of their position within the firm, managers not only enjoy valuable inside information concerning the firm's investment options and performance; they also enjoy valuable inside information concerning managerial performance. Outside detection of managerial shirking or self-dealing is made difficult by the fact that the same forces that provide incentives for chicanery or fraud in the management of a firm's assets provide incentives for self-dealing in the management of information. No manager can be reasonably expected to admit, "this firm is being robbed, and I'm the crook."

Of course, evidence of managerial self-dealing in information management would constitute only necessary but not sufficient evidence against the strong-form managerial labor market efficiency hypothesis. The functioning of the managerial labor market internal to the firm, and two-way managerial monitoring, has the theoretical potential to ensure efficiency in the face of asymmetric information between insiders and outsiders. However, the hierarchical structure of management within firms argues against this possibility. When managers within an organization are organized or classified according to rank, authority, or capacity ("talent"), the efficiency of internal two-way monitoring becomes distinctly one-sided. As we proceed from the bottom to the top of an organization structure, the level of managerial talent or effectiveness can be expected to rise. By

virtue of their past success in scaling the corporate ladder, top managers can be presumed relatively more efficient than lesser managers in the management of firm assets and information. "Survival of the fittest" argues for superior top-down vs. bottoms-up internal two-way monitoring, and therefore against the notion of perfect monitoring symmetry.

Figure 1 illustrates the managerial labor market implications of asymmetric information concerning managerial performance. For simplicity, a one-period model with perfectly inelastic supply is assumed. Take the lower $D_1 = MRP_1$ to be reflective of the "true" marginal revenue product of a manager, and therefore indicative of the economic value of managerial services. The effect of asymmetric information is to allow a greater and false "perceived" demand for managerial services, $D_2 = MRP_2$, to be created. The difference in managerial compensation $\Delta = P_2 - P_1$ constitutes a measure of the value to managers, and cost to employers (stockholders), resulting from managers' self-dealing in the dissemination of performance information.

A second main thesis of this analysis is that, just as in the case of asymmetric information concerning managerial performance, the presence of firm-specific human capital will create a wedge within the market demand curve for managers (see Robert Frank, 1984). In this instance, take the greater $D_2 = MRP_2$ to be reflective of the true marginal revenue product of a manager, and consisting of returns to both firm-specific as well as nonfirm-specific human capital. Take the lower $D_1 = MRP_1$ to reflect returns to nonfirm-specific human capital only. In a managerial labor market that is competitive in terms of supply, $D_1 = MRP_1$ illustrates the outside wage opportunity cost P_1 of managers; whereas $D_2 = MRP_2$ illustrates the maximum possible inside wage P_2, or economic value of managerial services. The difference in managerial compensation $\Delta = P_2 - P_1$ depicts a range within which bargaining outcomes in an imperfectly competitive managerial labor market will occur. Importantly, the indeterminacy of the managerial compensation bargain outcome which arises in the presence of firm-specific human capital

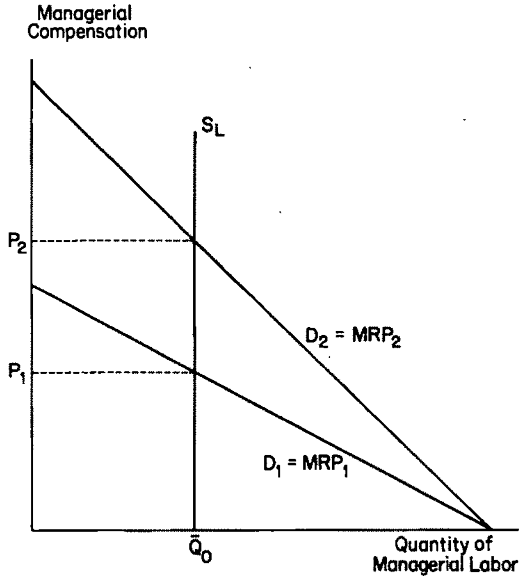


FIGURE 1. THE MARKET FOR MANAGERS

will by itself create an incentive for managerial self-dealing in the management of information. Therefore, asymmetric information and firm-specific human capital work singly and in concert to undermine the basis for strong-form managerial labor market efficiency.

Within this context, the market for corporate control can be seen as reflecting a wide variety of mechanisms for mitigating information agency costs. I suggest a semistrong-form of managerial labor market efficiency, namely: "Mergers, buyouts, and fakeouts are among a variety of mechanisms that complete the managerial labor market by efficiently minimizing the agency costs that arise due to the effects of asymmetric information and firm-specific human capital."

II. The Market for Corporate Control

A. *Friendly Mergers*

Jensen and Ruback, for example, offer superior efficiency in the management of acquired firm assets as an explanation of empirical evidence that target firm share-

holders benefit, and bidding firm shareholders do not lose in takeover "battles." Unfriendly takeovers can be especially unfriendly to inefficient management which is subsequently replaced. However, by focusing on mergers solely as a mechanism by which superior managerial teams replace inferior counterparts in acquired firms, the phenomenon of friendly mergers escapes explanation. I do not wish to dispute that, at least in part, merger activity reflects managerial competition for the right to manage corporate resources. Rather, I wish to suggest that unfriendly mergers, or buyouts, and friendly mergers are independent phenomena worthy of individual consideration.

Friendly mergers (including "white knight" acquisitions) can be defined as a change in corporate ownership without an accompanying change in managerial control. "Toe-hold" acquisitions, for example, are obviously made within the context of the *de novo* entry alternative. The calculus of the acquiring firm can be expected to include a comparative evaluation of its own and the target firm's human and nonhuman resources. When duplication of target firm capabilities or expertise becomes difficult and costly, acquisition as opposed to *de novo* entry can become attractive. Indeed, merger accompanied by retention of target firm management constitutes a revealed preference for acquisition rather than duplication of target firm managerial expertise. As such, it becomes prima facie evidence that the "buy" managerial expertise decision is more attractive, either less costly or less risky, than the "make," *de novo* entry, alternative.

The comparative attractiveness of the buy (merger) vs. make (entry) managerial expertise alternatives arises due to the unexploited firm-specific human capital of target firm management. In a perfectly competitive market for managerial talent, the human capital of target firm management would be fully priced and no differential would emerge in the buy vs. make alternatives. With firm-specific human capital, however, acquisition of target firm managerial expertise at less than its full economic value becomes possible. From the perspective of target firm management, friendly takeovers are means for obtaining increased returns to firm-specific human capital. Friendly mergers can therefore be considered valuable mechanisms which help complete the market for managers by driving the human capital rental rates of target firm management toward economic values.

B. *Unfriendly Buyouts*

An unfriendly buyout can be defined as a successful unsolicited takeover bid that results in the replacement of target firm management. As suggested earlier, this topic has been the worthy subject of substantial previous research. I am satisfied here to offer brief comments as to how asymmetric information and firm-specific human capital can be regarded as contributory explanations of the unfriendly buyout phenomenon.

Asymmetric information concerning target firm managerial performance provides an important motivation for unfriendly buyouts in that the capabilities that allow an individual to evaluate the skills of managers within a given organization, and thereby successfully climb the corporate ladder to the top, also lend themselves to the evaluation of competing management teams. Successful managers are especially well equipped to detect the failures of inefficient management, and to seize the opportunity provided by that inefficiency. It follows that the unfriendly buyout phenomenon will reflect, at least in part, the manager catalyst with insider (nonpublic) information concerning target firm managerial inefficiency.

In the presence of firm-specific human capital, acquiring firm management will also possess strong economic incentives to spread its expertise over broader economic resources. As the firm grows, management becomes able to spread its firm-specific human capital over greater related assets and thereby obtain larger, but still only partial, economic rentals. With growth, management also becomes able to enhance its tools and techniques through greater experience or learning. Growth in the scale and/or scope of the enterprise will cause the firm-specific compo-

nent of management human capital to shrink in relative terms. The nonfirm-specific component, and management opportunity cost, will rise. Therefore, unfriendly buyouts, especially conglomerate mergers, have the potential to enhance management's more marketable nonfirm-specific component of human capital and thereby provide a further motive for unfriendly buyouts.

C. *Fakeouts*

To this point, I have been both devious and unfair. While perhaps piquing your curiosity, I have briefly mentioned but not yet defined what I mean by fakeouts. At long last, I define fakeouts as that class of market mechanisms that allow management with little or no ownership interest to fend off unwanted suitors. In popular parlance these include "poison pill," "shark repellent," "greenmail," and other such defenses. Poison pill and shark repellent defenses describe corporate bylaws regarding election of the board of directors, setting the date of annual meetings, establishing share purchase minimums for hostile bids (for example, 100 percent of outstanding shares) and other such requirements that make successful hostile takeovers difficult. Greenmail describes a payment in the form of a minority interest buyout above acquisition costs in return for a "raiders" agreement to terminate a hostile takeover bid.

Poison pill and shark repellent defenses can have the *ex ante*, or before takeover bid, effect of depressing the market value of target firms by reducing the probability of a successful hostile takeover. Conversely, the expectation of receiving greenmail payments can increase the *ex ante* market value of target firms by raising raiders' expected return on a hostile bid. The *ex post*, or after successful takeover defense, effect on target firm market values for each type of defense can be expected to be negative. Perhaps it is this negative *ex post* effect on target firm market values that is responsible for the calls for regulation that have arisen in Congress and elsewhere (see Gregg Jarrell and Michael Bradley, 1980). Why should "entrenched" target firm management benefit at the expense of target firm stockholders?

I suggest that the perceived benefit to target firm management, and cost to target firm stockholders, resulting from a successful hostile takeover defense is more apparent than real. The economic cost to target firm stockholders of any successful takeover defense is constrained by the market for managerial talent to be no more than the discounted present value of target firm management's uncompensated firm-specific human capital. When the cost of takeover defense covenants falls short of the unexploited value of management's firm-specific human capital, they will be adopted by stockholders who recognize the residual net benefits to retaining incumbent management. When the cost of takeover defense covenants exceeds the value of management's unexploited firm-specific human capital, they will be rejected by stockholders who recognize the residual net costs to retaining incumbent management. Therefore, fakeouts can be thought of as market mechanisms that permit incumbent management to obtain a greater share of the rentals accruing to their firm-specific human capital. As such, fakeouts complement friendly mergers and hostile buyouts as mechanisms in the market for corporate control which help complete the market for managerial talent.

III. Conclusion

This paper argues that friendly mergers, hostile takeover bids, and fakeouts (including a variety of takeover defenses) can be considered as market mechanisms that help complete the market for managerial talent. In the presence of asymmetric information concerning managerial performance and firm-specific human capital, a semistrong form of managerial labor market efficiency is suggested. These mechanisms serve to minimize agency costs related to managerial self-dealing in information management; while at the same time providing a means through which the rental rates of managers can more fully reflect full economic values, including the returns to firm-specific human capital.

Therefore, imperfections in the market for managerial talent create a kind of jointness between the market for managers and the market for corporate control.

REFERENCES

Berle, Adolf A. and Means, Gardiner C., *The Modern Corporation and Private Property*, New York: Macmillan, 1932.

Fama, Eugene F., "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, May 1970, *25*, 597–610.

______, *Foundations of Finance*, New York: Basic Books, 1976.

______, "Agency Problems and the Theory of the Firm," *Journal of Political Economy*, April 1980, *88*, 288–307.

Frank, Robert H., "Are Workers Paid Their Marginal Products?," *American Economic Review*, September 1984, *74*, 549–71.

Jarrell, Gregg A. and Bradley, Michael, "The Economic Effects of Federal and State Regulations of Cash Tender Offers," *Journal of Law and Economics*, October 1980, *23*, 371–407.

Jensen, Michael C. (ed.), "Symposium on the Market for Corporate Control," *Journal of Financial Economics*, April 1983, *11*, 1–475.

______ **and Meckling, William H.,** "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, October 1976, *3*, 305–60.

______ **and Ruback, Richard S.,** "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics*, April 1983, *11*, 5–50.

______ **and Zimmerman, Jerold L.** (eds.), "Symposium on Management Compensation and the Managerial Labor Market in Honor of William H. Meckling," *Journal of Accounting and Economics*, April 1985, *7*, 1–257.

Manne, Henry G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, April 1965, *73*, 110–20.

Moore, Thomas Gale (ed.), "Corporations and Private Property: A Conference Sponsored by the Hoover Institution," *Journal of Law and Economics*, June 1983, *26*, 235–496.

Schwert, G. William, "Size and Stock Returns, and Other Empirical Regularities," *Journal of Financial Economics*, June 1983, *12*, 3–12.

Smith, Adam, *The Wealth of Nations* (Glasgow ed. 1976), 1776.

Williamson, Oliver, "Managerial Discretion and Business Behavior," *American Economic Review*, December 1963, *53*, 1032–57.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington: USGPO, 1985, 187–216.

Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers

By MICHAEL C. JENSEN*

Corporate managers are the agents of shareholders, a relationship fraught with conflicting interests. Agency theory, the analysis of such conflicts, is now a major part of the economics literature. The payout of cash to shareholders creates major conflicts that have received little attention.[1] Payouts to shareholders reduce the resources under managers' control, thereby reducing managers' power, and making it more likely they will incur the monitoring of the capital markets which occurs when the firm must obtain new capital (see M. Rozeff, 1982; F. H. Easterbrook, 1984). Financing projects internally avoids this monitoring and the possibility the funds will be unavailable or available only at high explicit prices.

Managers have incentives to cause their firms to grow beyond the optimal size. Growth increases managers' power by increasing the resources under their control. It is also associated with increases in managers' compensation, because changes in compensation are positively related to the growth in sales (see Kevin Murphy, 1985). The tendency of firms to reward middle managers through promotion rather than year-to-year bonuses also creates a strong organizational bias toward growth to supply the new positions that such promotion-based reward systems require (see George Baker, 1986).

Competition in the product and factor markets tends to drive prices towards minimum average cost in an activity. Managers must therefore motivate their organizations to increase efficiency to enhance the probability of survival. However, product and factor market disciplinary forces are often weaker in new activities and activities that involve substantial economic rents or quasi rents.[2] In these cases, monitoring by the firm's internal control system and the market for corporate control are more important. Activities generating substantial economic rents or quasi rents are the types of activities that generate substantial amounts of free cash flow.

Free cash flow is cash flow in excess of that required to fund all projects that have positive net present values when discounted at the relevant cost of capital. Conflicts of interest between shareholders and managers over payout policies are especially severe when the organization generates substantial free cash flow. The problem is how to motivate managers to disgorge the cash rather than investing it at below the cost of capital or wasting it on organization inefficiencies.

The theory developed here explains 1) the benefits of debt in reducing agency costs of free cash flows, 2) how debt can substitute

*LaClare Professor of Finance and Business Administration and Director of the Managerial Economics Research Center, University of Rochester Graduate School of Management, Rochester, NY 14627, and Professor of Business Administration, Harvard Business School. This research is supported by the Division of Research, Harvard Business School, and the Managerial Economics Research Center, University of Rochester. I have benefited from discussions with George Baker, Gordon Donaldson, Allen Jacobs, Jay Light, Clifford Smith, Wolf Weinhold, and especially Armen Alchian and Richard Ruback.

[1] Gordon Donaldson (1984) in his study of 12 large *Fortune* 500 firms concludes that managers of these firms were not driven by maximization of the value of the firm, but rather by the maximization of "corporate wealth," defined as "*the aggregate purchasing power available to management for strategic purposes during any given planning period*" (p. 3). "In practical terms it is cash, credit, and other corporate purchasing power by which management commands goods and services" (p. 22).

[2] Rents are returns in excess of the opportunity cost of the resources to the activity. Quasi rents are returns in excess of the short-run opportunity cost of the resources to the activity.

for dividends, 3) why "diversification" programs are more likely to generate losses than takeovers or expansion in the same line of business or liquidation-motivated takeovers, 4) why the factors generating takeover activity in such diverse activities as broadcasting and tobacco are similar to those in oil, and 5) why bidders and some targets tend to perform abnormally well prior to takeover.

I. The Role of Debt in Motivating Organizational Efficiency

The agency costs of debt have been widely discussed, but the benefits of debt in motivating managers and their organizations to be efficient have been ignored. I call these effects the "control hypothesis" for debt creation.

Managers with substantial free cash flow can increase dividends or repurchase stock and thereby pay out current cash that would otherwise be invested in low-return projects or wasted. This leaves managers with control over the use of future free cash flows, but they can promise to pay out future cash flows by announcing a "permanent" increase in the dividend. Such promises are weak because dividends can be reduced in the future. The fact that capital markets punish dividend cuts with large stock price reductions is consistent with the agency costs of free cash flow.

Debt creation, without retention of the proceeds of the issue, enables managers to effectively bond their promise to pay out future cash flows. Thus, debt can be an effective substitute for dividends, something not generally recognized in the corporate finance literature. By issuing debt in exchange for stock, managers are bonding their promise to pay out future cash flows in a way that cannot be accomplished by simple dividend increases. In doing so, they give shareholder recipients of the debt the right to take the firm into bankruptcy court if they do not maintain their promise to make the interest and principle payments. Thus debt reduces the agency costs of free cash flow by reducing the cash flow available for spending at the discretion of managers. These control effects of debt are a potential determinant of capital structure.

Issuing large amounts of debt to buy back stock also sets up the required organizational incentives to motivate managers and to help them overcome normal organizational resistance to retrenchment which the payout of free cash flow often requires. The threat caused by failure to make debt service payments serves as an effective motivating force to make such organizations more efficient. Stock repurchase for debt or cash also has tax advantages. (Interest payments are tax deductible to the corporation, and that part of the repurchase proceeds equal to the seller's tax basis in the stock is not taxed at all.)

Increased leverage also has costs. As leverage increases, the usual agency costs of debt rise, including bankruptcy costs. The optimal debt-equity ratio is the point at which firm value is maximized, the point where the marginal costs of debt just offset the marginal benefits.

The control hypothesis does not imply that debt issues will always have positive control effects. For example, these effects will not be as important for rapidly growing organizations with large and highly profitable investment projects but no free cash flow. Such organizations will have to go regularly to the financial markets to obtain capital. At these times the markets have an opportunity to evaluate the company, its management, and its proposed projects. Investment bankers and analysts play an important role in this monitoring, and the market's assessment is made evident by the price investors pay for the financial claims.

The control function of debt is more important in organizations that generate large cash flows but have low growth prospects, and even more important in organizations that must shrink. In these organizations the pressures to waste cash flows by investing them in uneconomic projects is most serious.

II. Evidence from Financial Restructuring

The free cash flow theory of capital structure helps explain previously puzzling results

on the effects of financial restructuring. My paper with Clifford Smith (1985, Table 2) and Smith (1986, Tables 1 and 3) summarize more than a dozen studies of stock price changes at announcements of transactions which change capital structure. Most leverage-increasing transactions, including stock repurchases and exchange of debt or preferred for common, debt for preferred, and income bonds for preferred, result in significantly positive increases in common stock prices. The 2-day gains range from 21.9 percent (debt for common) to 2.2 percent (debt or income bonds for preferred). Most leverage-reducing transactions, including the sale of common, and exchange of common for debt or preferred, or preferred for debt, and the call of convertible bonds or convertible preferred forcing conversion into common, result in significant decreases in stock prices. The 2-day losses range from -9.9 percent (common for debt) to $-.4$ percent (for call of convertible preferred forcing conversion to common). Consistent with this, free cash flow theory predicts that, except for firms with profitable unfunded investment projects, prices will rise with unexpected increases in payouts to shareholders (or promises to do so), and prices will fall with reductions in payments or new requests for funds (or reductions in promises to make future payments).

The exceptions to the simple leverage change rule are targeted repurchases and the sale of debt (of all kinds) and preferred stock. These are associated with abnormal price declines (some of which are insignificant). The targeted repurchase price decline seems to be due to the reduced probability of takeover. The price decline on the sale of debt and preferred stock is consistent with the free cash flow theory because these sales bring new cash under the control of managers. Moreover, the magnitudes of the value changes are positively related to the change in the tightness of the commitment bonding the payment of future cash flows, for example, the effects of debt for preferred exchanges are smaller than the effects of debt for common exchanges. Tax effects can explain some of these results, but not all, for example, the price increases on exchange of preferred for common, which has no tax effects.

III. Evidence from Leveraged Buyout and Going Private Transactions

Many of the benefits in going private and leveraged buyout (*LBO*) transactions seem to be due to the control function of debt. These transactions are creating a new organizational form that competes successfully with the open corporate form because of advantages in controlling the agency costs of free cash flow. In 1984, going private transactions totaled $10.8 billion and represented 27 percent of all public acquisitions (by number, see W. T. Grimm, 1985, Figs. 36 and 37). The evidence indicates premiums paid average over 50 percent.[3]

Desirable leveraged buyout candidates are frequently firms or divisions of larger firms that have stable business histories and substantial free cash flow (i.e., low growth prospects and high potential for generating cash flows)—situations where agency costs of free cash flow are likely to be high. The *LBO* transactions are frequently financed with high debt; 10 to 1 ratios of debt to equity are not uncommon. Moreover, the use of strip financing and the allocation of equity in the deals reveal a sensitivity to incentives, conflicts of interest, and bankruptcy costs.

Strip financing, the practice in which risky nonequity securities are held in approximately equal proportions, limits the conflict of interest among such securities' holders and therefore limits bankruptcy costs. A somewhat oversimplified example illustrates the point. Consider two firms identical in every respect except financing. Firm A is entirely financed with equity, and firm B is highly leveraged with senior subordinated debt, convertible debt and preferred as well

[3]See H. DeAngelo et al. (1984), and L. Lowenstein (1985). Lowenstein also mentions incentive effects of debt, but argues tax effects play a major role in explaining the value increase.

as equity. Suppose firm B securities are sold only in strips, that is, a buyer purchasing X percent of any security must purchase X percent of all securities, and the securities are "stapled" together so they cannot be separated later. Security holders of both firms have identical unlevered claims on the cash flow distribution, but organizationally the two firms are very different. If firm B managers withhold dividends to invest in value-reducing projects or if they are incompetent, strip holders have recourse to remedial powers not available to the equity holders of firm A. Each firm B security specifies the rights its holder has in the event of default on its dividend or coupon payment, for example, the right to take the firm into bankruptcy or to have board representation. As each security above the equity goes into default, the strip holder receives new rights to intercede in the organization. As a result, it is easier and quicker to replace managers in firm B.

Moreover, because every security holder in the highly levered firm B has the same claim on the firm, there are no conflicts among senior and junior claimants over reorganization of the claims in the event of default; to the strip holder it is a matter of moving funds from one pocket to another. Thus firm B need never go into bankruptcy, the reorganization can be accomplished voluntarily, quickly, and with less expense and disruption than through bankruptcy proceedings.

Strictly proportional holdings of all securities is not desirable, for example, because of IRS restrictions that deny tax deductibility of debt interest in such situations and limits on bank holdings of equity. However, riskless senior debt needn't be in the strip, and it is advantageous to have top-level managers and venture capitalists who promote the transactions hold a larger share of the equity. Securities commonly subject to strip practices are often called "mezzanine" financing and include securities with priority superior to common stock yet subordinate to senior debt.

Top-level managers frequently receive 15–20 percent of the equity. Venture capitalists and the funds they represent retain the major share of the equity. They control the board of directors and monitor managers. Managers and venture capitalists have a strong interest in making the venture successful because their equity interests are subordinate to other claims. Success requires (among other things) implementation of changes to avoid investment in low return projects to generate the cash for debt service and to increase the value of equity. Less than a handful of these ventures have ended in bankruptcy, although more have gone through private reorganizations. A thorough test of this organizational form requires the passage of time and another recession.

IV. Evidence from the Oil Industry

Radical changes in the energy market since 1973 simultaneously generated large increases in free cash flow in the petroleum industry and required a major shrinking of the industry. In this environment the agency costs of free cash flow were large, and the takeover market has played a critical role in reducing them. From 1973 to the late 1970's, crude oil prices increased tenfold. They were initially accompanied by increases in expected future oil prices and an expansion of the industry. As consumption of oil fell, expectations of future increases in oil prices fell. Real interest rates and exploration and development costs also increased. As a result the optimal level of refining and distribution capacity and crude reserves fell in the late 1970's and early 1980's, leaving the industry with excess capacity. At the same time profits were high. This occurred because the average productivity of resources in the industry increased while the marginal productivity decreased. Thus, contrary to popular beliefs, the industry had to shrink. In particular, crude oil reserves (the industry's major asset) were too high, and cutbacks in exploration and development ($E\&D$) expenditures were required (see my 1986 paper).

Price increases generated large cash flows in the industry. For example, 1984 cash flows of the ten largest oil companies were \$48.5 billion, 28 percent of the total cash flows of

the top 200 firms in Dun's *Business Month* survey. Consistent with the agency costs of free cash flow, management did not pay out the excess resources to shareholders. Instead, the industry continued to spend heavily on *E&D* activity even though average returns were below the cost of capital.

Oil industry managers also launched diversification programs to invest funds outside the industry. The programs involved purchases of companies in retailing (Marcor by Mobil), manufacturing (Reliance Electric by Exxon), office equipment (Vydec by Exxon), and mining (Kennecott by Sohio, Anaconda Minerals by Arco, Cyprus Mines by Amoco). These acquisitions turned out to be among the least successful of the last decade, partly because of bad luck (for example, the collapse of the minerals industry) and partly because of a lack of managerial expertise outside the oil industry. Although acquiring firm shareholders lost on these acquisitions, the purchases generated social benefits to the extent they diverted cash to shareholders (albeit to target shareholders) that otherwise would have been wasted on unprofitable real investment projects.

Two studies indicate that oil industry exploration and development expenditures have been too high since the late 1970's. John McConnell and Chris Muscarella (1986) find that announcements of increases in *E&D* expenditures by oil companies in the period 1975–81 were associated with systematic *decreases* in the announcing firm's stock price, and vice versa. These results are striking in comparison with their evidence that the opposite market reaction occurs to changes in investment expenditures by industrial firms, and similar SEC evidence on increases in *R&D* expenditures. (See Office of the Chief Economist, SEC, 1985.) B. Picchi's study of returns on *E&D* expenditures for 30 large oil firms indicates on average the industry did not earn "...even a 10% return on its pretax outlays" (1985, p. 5) in the period 1982–84. Estimates of the average ratio of the present value of future net cash flows of discoveries, extensions, and enhanced recovery to *E&D* expenditures for the industry ranged from less than 60 to 90 cents on every dollar invested in these activities.

V. Takeovers in the Oil Industry

Retrenchment requires cancellation or delay of many ongoing and planned projects. This threatens the careers of the people involved, and the resulting resistance means such changes frequently do not get made in the absence of a crisis. Takeover attempts can generate crises that bring about action where none would otherwise occur.

Partly as a result of Mesa Petroleum's efforts to extend the use of royalty trusts which reduce taxes and pass cash flows directly through to shareholders, firms in the oil industry were led to merge, and in the merging process they incurred large increases in debt, paid out large amounts of capital to shareholders, reduced excess expenditures on *E&D* and reduced excess capacity in refining and distribution. The result has been large gains in efficiency and in value. Total gains to shareholders in the Gulf/Chevron, Getty/Texaco, and Dupont/Conoco mergers, for example, were over \$17 billion. More is possible. Allen Jacobs (1986) estimates total potential gains of about \$200 billion from eliminating inefficiencies in 98 firms with significant oil reserves as of December 1984.

Actual takeover is not necessary to induce the required retrenchment and return of resources to shareholders. The restructuring of Phillips and Unocal (brought about by threat of takeover) and the voluntary Arco restructuring resulted in stockholder gains ranging from 20 to 35 percent of market value (totaling \$6.6 billion). The restructuring involved repurchase of from 25 to 53 percent of equity (for over \$4 billion in each case), substantially increased cash dividends, sales of assets, and major cutbacks in capital spending (including *E&D* expenditures). Diamond-Shamrock's reorganization is further support for the theory because its market value *fell* 2 percent on the announcement day. Its restructuring involved, among other things, *reducing* cash dividends by 43 percent, repurchasing 6 percent of its shares for \$200 million, selling 12 percent of a newly created

master limited partnership to the public, and *increasing* expenditures on oil and gas exploration by $100 million/year.

VI. Free Cash Flow Theory of Takeovers

Free cash flow is only one of approximately a dozen theories to explain takeovers, all of which I believe are of some relevance (see my 1986 paper). Here I sketch out some empirical predictions of the free cash flow theory, and what I believe are the facts that lend it credence.

The positive market response to debt creation in oil industry takeovers (as well as elsewhere, see Robert Bruner, 1985) is consistent with the notion that additional debt increases efficiency by forcing organizations with large cash flows but few high-return investment projects to disgorge cash to investors. The debt helps prevent such firms from wasting resources on low-return projects.

Free cash flow theory predicts which mergers and takeovers are more likely to destroy, rather than to create, value; it shows how takeovers are both evidence of the conflicts of interest between shareholders and managers, and a solution to the problem. Acquisitions are one way managers spend cash instead of paying it out to shareholders. Therefore, the theory implies managers of firms with unused borrowing power and large free cash flows are more likely to undertake low-benefit or even value-destroying mergers. Diversification programs generally fit this category, and the theory predicts they will generate lower total gains. The major benefit of such transactions may be that they involve less waste of resources than if the funds had been internally invested in unprofitable projects. Acquisitions not made with stock involve payout of resources to (target) shareholders and this can create net benefits even if the merger generates operating inefficiencies. Such low-return mergers are more likely in industries with large cash flows whose economics dictate that exit occur. In declining industries, mergers within the industry will create value, and mergers outside the industry are more likely to be low- or even negative-return projects. Oil fits this description and so does tobacco. Tobacco firms face declining demand due to changing smoking habits but generate large free cash flow and have been involved in major acquisitions recently. Forest products is another industry with excess capacity. Food industry mergers also appear to reflect the expenditure of free cash flow. The industry apparently generates large cash flows with few growth opportunities. It is therefore a good candidate for leveraged buyouts and these are now occurring. The $6.3 billion Beatrice *LBO* is the largest ever. The broadcasting industry generates rents in the form of large cash flows on its licenses and also fits the theory. Regulation limits the supply of licenses and the number owned by a single entity. Thus, profitable internal investments are limited and the industry's free cash flow has been spent on organizational inefficiencies and diversification programs—making these firms takeover targets. CBS's debt for stock restructuring fits the theory.

The theory predicts value increasing takeovers occur in response to breakdowns of internal control processes in firms with substantial free cash flow and organizational policies (including diversification programs) that are wasting resources. It predicts hostile takeovers, large increases in leverage, dismantlement of empires with few economies of scale or scope to give them economic purpose (for example, conglomerates), and much controversy as current managers object to loss of their jobs or the changes in organizational policies forced on them by threat of takeover.

The debt created in a hostile takeover (or takeover defense) of a firm suffering severe agency costs of free cash flow is often not permanent. In these situations, levering the firm so highly that it cannot continue to exist in its old form generates benefits. It creates the crisis to motivate cuts in expansion programs and the sale of those divisions which are more valuable outside the firm. The proceeds are used to reduce debt to a more normal or permanent level. This process results in a complete rethinking of the organization's strategy and its structure. When successful a much leaner and competitive

organization results.

Consistent with the data, free cash flow theory predicts that many acquirers will tend to have exceptionally good performance prior to acquisition. (Again, the oil industry fits well.) That exceptional performance generates the free cash flow for the acquisition. Targets will be of two kinds: firms with poor management that have done poorly prior to the merger, and firms that have done exceptionally well and have large free cash flow which they refuse to pay out to shareholders. Both kinds of targets seem to exist, but more careful analysis is desirable (see D. Mueller, 1980).

The theory predicts that takeovers financed with cash and debt will generate larger benefits than those accomplished through exchange of stock. Stock acquisitions tend to be different from debt or cash acquisitions and more likely to be associated with growth opportunities and a shortage of free cash flow; but that is a topic for future consideration.

The agency cost of free cash flow is consistent with a wide range of data for which there has been no consistent explanation. I have found no data which is inconsistent with the theory, but it is rich in predictions which are yet to be tested.

REFERENCES

Baker, George, "Compensation and Hierarchies," Harvard Business School, January 1986.

Bruner, Robert F., "The Use of Excess Cash and Debt Capacity as a Motive for Merger," Colgate Darden Graduate School of Business, December 1985.

DeAngelo, H., DeAngelo, L. and Rice, E., "Going Private: Minority Freezeouts and Stockholder Wealth," *Journal of Law and Economics*, October 1984, *27*, 367–401.

Donaldson, Gordon, *Managing Corporate Wealth*, New York: Praeger, 1984.

Easterbrook, F. H., "Two Agency-Cost Explanations of Dividends," *American Economic Review*, September 1984, *74*, 650–59.

Grimm, W. T., *Mergerstat Review*, 1985.

Jacobs, E. Allen, "The Agency Cost of Corporate Control," MIT, February 6, 1986.

Jensen, Michael C., "The Takeover Controversy: Analysis and Evidence," Managerial Economics Research Center, Working Paper No. 86-01, University of Rochester, March 1986.

______, "When Unocal Won Over Pickens, Shareholders and Society Lost," *Financier*, November 1985, *9*, 50–52.

______ **and Smith, C. W., Jr.,** "Stockholder, Manager and Creditor Interests: Applications of Agency Theory," in E. Altman and M. Subrahmanyam, eds., *Recent Advances in Corporate Finance*, Homewood: Richard Irwin, 1985, 93–131.

Lowenstein, L., "Management Buyouts," *Columbia Law Review*, May 1985, *85*, 730–84.

McConnell, John J. and Muscarella, Chris J., "Corporate Capital Expenditure Decisions and the Market Value of the Firm," *Journal of Financial Economics*, forthcoming 1986.

Mueller, D., *The Determinants and Effects of Mergers*, Cambridge: Oelgeschlager, 1980.

Murphy, Kevin J., "Corporate Performance and Managerial Remuneration: An Empirical Analysis," *Journal of Accounting and Economics*, April 1985, *7*, 11–42.

Picchi, B., "Structure of the U.S. Oil Industry: Past and Future," Salomon Brothers, July 1985.

Rozeff, M., "Growth, Beta and Agency Costs as Determinants of Dividend Payout Ratios," *Journal of Financial Research*, Fall 1982, *5*, 249–59.

Smith, Clifford W., "Investment Banking and the Capital Acquisition Process," *Journal of Financial Economics*, Nos. 1–2, *15*, forthcoming, 1986.

***Dun's Business Month*,** "Cash Flow: The Top 200," July 1985, 44–50.

Office of the Chief Economist, SEC, "Institutional Ownership, Tender Offers, and Long-Term Investments," April 1985.

The International Transmission and Effects of Fiscal Policies

By JACOB A. FRENKEL AND ASSAF RAZIN*

In recent years the world economy has been subject to large and unsynchronized changes in fiscal policies, high and volatile real rates of interest, large fluctuations in real exchange rates, and significant variations in private-sector spending. During the first half of the 1980's national fiscal policies have exhibited large divergencies. The United States adopted an expansionary course while the other major countries taken together followed a relatively contractionary course. Policies undertaken by the major economies affected the rest of the world through the integrated capital market. This paper deals with the international transmission of fiscal policies and their effects on real exchange rates and real interest rates. Section I reviews key facts and Section II provides an analytical framework relevant for the interpretation of these facts.

I. Selected Facts

Since the beginning of 1980, short- and long-term real rates of interest exhibited different patterns. A weighted average of the annual short-term real interest rates in the five major industrial countries (the United States, Canada, Japan, Germany, and the United Kingdom) rose from 2.1 percent in January 1980 to 4.0 percent in July 1985; the corresponding long-term rates rose from 0.6 percent in January 1980 to 5.7 percent in July 1985. Both rates peaked and surpassed 8 percent in mid-1982. Thus during 1980–85, real rates of interest have been high (in comparison with early 1980) and the slope of the real yield curve which was negative until the third quarter of 1981, has turned positive starting from mid-1982. The same period also witnessed sharp changes in real exchange rates. In the first quarter of 1985, the real effective value of the U.S. dollar was about 43 percent above its average value for the decade 1974–83 and 57 percent above its low point of the third quarter of 1980. (The source of all data used in this paper is IMF, *World Economic Outlook*, 1985.)

These changes in real interest rates and real exchange rates were associated with large and divergent changes in world fiscal policies. The budget deficit of the general U.S. government as a fraction of *GNP* rose from about 1 percent in 1980 to about 3.5 percent in 1985 (after reaching a peak of 4.1 percent in 1983). At the same time, the budget deficit as a fraction of *GNP* declined in Japan, Germany, and the United Kingdom. Similarly, since 1980 according to IMF measures, the fiscal impulse (which is a more exogenous measure of fiscal policy) has been expansionary for the United States and contractionary for the other major industrial countries taken together. Another indicator of the levels and divergence among national fiscal policies is provided by a comparison among annual percentage changes in public-sector consumption. As seen in Table 1, the percentage annual growth of U.S. public-sector consumption accelerated in the past 2 years exceeding 4 percent in 1985. During the late 1970's and early 1980's, public-sector consumption in Japan grew faster than in the United States (the difference reaching 4.3 percent in 1981), and during the past 2 years, it grew more slowly (the difference in "favor"

†*Discussants*: Robert J. Gordon, Northwestern University; William H. Branson, Princeton University.

*Professors of Economics, University of Chicago, Chicago, IL 60637, and the NBER; Tel Aviv University, Tel Aviv, Israel and the NBER, respectively.

TABLE 1—DIFFERENCES IN PRIVATE AND PUBLIC CONSUMPTION: UNITED STATES AND JAPAN 1977–85[a]

	U.S.		U.S. minus Japan	
	Private	Public	Private	Public
1977	5.0	1.5	+1.2	−2.4
1978	4.5	2.0	−0.2	−3.1
1979	2.7	1.3	−3.2	−3.0
1980	0.5	2.2	−0.8	−0.7
1981	2.0	0.9	+1.2	−4.3
1982	1.4	2.0	−2.8	+0.2
1983	4.8	−0.3	+1.5	−3.2
1984	5.3	3.5	+2.4	+1.2
1985	4.1	4.5	+0.7	+2.1

Source: IMF *World Economic Outlook*, October 1985.
[a]Annual percentage changes.

of the United States reaching 2.1 percent in 1985).

Concomitantly, the annual percentage changes in real private-sector consumption also displayed large fluctuations that differed across countries. In the United States these changes ranged from 0.5 percent in 1980 to 5.3 percent in 1984 and, as seen in Table 1, the growth of private-sector consumption in Japan exceeded that in the United States during 1978–80 and fell short of it during 1983–85 (the differential growth rate of fixed investment displays a similar pattern).

II. A Conceptual Framework

In this section we outline a simple two-country model of the world economy suitable for an interpretation of the facts outlined in Section I. The model provides insights into the interactions among fiscal policies, interest rates, real exchange rates, and the comovements of private-sector consumption. In order to deal with the real exchange rate, we assume that each country produces internationally tradable and nontradable goods and, in view of the high correlation among national real rates of interest, we focus on *world* rates of interest and assume that individuals have unlimited access to perfect world capital markets and that there are no distortions. For a meaningful analysis of budget deficits, we depart from the "Ricardian proposition" and introduce a "myopic" element as in Olivier Blanchard (1985). Accordingly, there are overlapping generations of rational individuals, but due to mortality each individual has a finite horizon. The coefficient of "myopia" reflects the finiteness of the horizon. Suppose that γ is the probability that an individual survives from one period to the next and let $\gamma<1$. The magnitude of γ influences savings in two ways. First, it introduces a risk premium $(1-\gamma)$ that raises the rate of interest applicable to individuals, ρ, above the world rate of interest, r, where $\rho=r+(1-\gamma)$. Hence, it impacts on current wealth through the heavier discounting of future disposable incomes. Second, it lowers the *effective* saving propensity from δ (in the absence of mortality) to $\gamma\delta$.

Government budgets are intertemporally balanced and government commitments are honored. Hence, government debt (at the beginning of period zero) equals the present values of current and future budget surpluses, and the discount rate applicable to government debt is the world rate of interest, r. We first divide the horizon into the current period and the future period. All quantities pertaining to the current period are indicated by a zero subscript and the paths of the exogenous variables are assumed stationary across future periods.

Equilibrium necessitates that in the current period, world output of tradable goods is demanded and the discounted sum of future outputs of tradable goods equals the discounted sums of future domestic and foreign demands. Likewise, in each country, current and future period outputs of nontradable goods must be demanded. The conditions that world markets for tradable goods clear in both current and future periods are stated in equations (1) and (2). These equations already incorporate the requirement that in each country the markets for nontradable goods clear.

$$(1)\quad (1-\beta)(1-\gamma\delta)W_0+(1-\beta^*)(1-\gamma\delta^*)\times W_0^*\left[r, B_{g0}^*; B_0\right]=\bar{Y}_{T0}-\bar{G}_{T0}$$

$$(2)\quad (1-\beta)\left(\gamma\delta W_0 + ((1-\gamma)(1+r))/r\rho \times I[r, W_0; T, \theta]\right) + (1-\beta^*)\left(\gamma\delta^* W_0^*[r, B_{g0}^*; B_0] + [((1-\gamma)(1+r))/r\rho] I^*[r, B_{g0}^*; B_0]\right) = \frac{1}{r}(\bar{Y}_T - \bar{G}_T)$$

where $\rho = r + 1 - \gamma$, and

$$B_{g0}^* = T_0^* - G_{T0}^* - \theta_0^* Y_{N0}^* p_0^*[r, B_{g0}^*; B_0] + (1/r)\left(T^* - G_T^* - \theta^* Y_N^* p^*[r, B_{g0}^*; B_0]\right).$$

Equation (1) states that the sum of world private demand for current tradable goods equals world supply ($\bar{Y}_{T0} = Y_{T0} + Y_{T0}^*$) net of government spending ($\bar{G}_{T0} = G_{T0} + G_{T0}^*$). In equation (1), W_0 denotes aggregate domestic private wealth in period zero, $(1-\gamma\delta)$ denotes the spending propensity, and $(1-\beta)$ is the consumption share of tradable goods. Hence, $(1-\beta)(1-\gamma\delta)W_0$ is the home country's private demand. Analogously, foreign private demand is $(1-\beta^*)(1-\gamma\delta^*)$ times foreign wealth W_0^*. The specification of private demands as a function of aggregate wealth reflects the assumption that individuals have an unlimited access to world capital markets. The value of wealth equals the difference between the discounted sum of labor income and net private debt. In equation (1), foreign current wealth W_0^* is expressed as a negative function of the rate of interest, reflecting the role of r in discounting future incomes, and as a negative function of private debt B_{p0}^* (square brackets indicate functional dependence). The latter in turn equals the difference between the foreign country's net external debt (which is the negative of the home country's net external debt, $-B_0$) and its net government debt, B_{g0}^*.

Equation (2) states that the discounted sum of domestic and foreign demands for future tradable goods equals the discounted sum of future world supply net of government spending. The first term is the product of the consumption share of tradable goods $(1-\beta)$ and total domestic future consumption. The latter equals the sum of the savings of those alive in period zero, $\gamma\delta W_0$, and the discounted sum of the demand for future goods of those who will be born in the future and whose disposable income in each period is I.[1] Disposable income (in terms of tradable goods) depends negatively on taxes, T, and positively on the relative price of nontradable goods p which in turn depends negatively on r (through its effect on future wealth of those yet unborn) and positively on W_0 (through its effect on the demand of those alive). The price, and thereby disposable income, also depends positively on the parameter θ which measures the share of output of nontradable goods absorbed by the government. Analogous interpretation applies to the second term on the left-hand side of equation (2) representing the foreign demand for future tradable goods. In specifying foreign disposable income, I^*, we incorporated the functional dependence of W_0^* on B_{g0}^* and B_0. The right-hand side of equation (2) denotes the discounted sum of world supply of future tradable goods net of government spending. Finally, the explicit expression for B_{g0}^* reflects the intertemporal budget constraint of the foreign government by which initial government debt must equal the discounted sum of current and future budget surpluses. In that expression, the terms $\theta_0^* Y_{N0}^* p_0^*$ and $\theta^* Y_N^* p^*$ measure current and future foreign government spending on nontradable goods, Y_N, where p^* is ex-

[1] In order to verify this, we note that in this model $1/(1-\gamma)$ is the population size and hence $(1-\gamma)I$ is each cohort's disposable income. Since the effective discount factor is $\gamma/(1+r)$, the wealth of each cohort is the discounted sum of each cohort's disposable income $[(1-\gamma)I]/[\rho/(1+r)]$ where $(1+r)/\rho$ is the annuity value of a perpetuity discounted by the effective discount factor. Since in each period there is a newly born cohort, the discounted sum of all cohorts incomes is $(1/r)$ times each cohort's wealth. For derivations of equations (1)–(2), see our article (1986b).

pressed as a negative function of r and a positive function of B_{g0}^{*} and B_0.

Equations (1) and (2) yield the equilibrium values of the home-country's initial wealth W_0, and the world rate of interest r, for any given values of the parameters. In equilibrium the demand for nontradable goods $\beta(1-\gamma\delta)W_0$ equals the supply net of government absorption $(1-\theta_0)p_0Y_{N0}$. Hence, the equilibrium price (the inverse of the real exchange rate) is $p_0=\beta(1-\gamma\delta)W_0/[(1-\theta_0)Y_{N0}]$. The equilibrium is represented by point A in Figure 1. The PP schedule shows combinations of r and p_0 that clear the market for present tradable goods. It is positively sloped since a rise in r lowers foreign demand (by lowering W_0^*) and a rise in p_0 raises domestic demand (by raising W_0). Future tradable goods market clears along the FF schedule. For a relatively small nontradable goods sector, the FF schedule is negatively sloped, since a rise in r creates an excess demand for future tradable goods which must be offset by a fall in W_0 (and therefore p_0).

A budget deficit arising from a current tax cut necessitates a corresponding rise in future taxes. As seen from equation (2) the rise in future taxes lowers domestic disposable income, I, and lowers the demand for future goods. For a given world rate of interest, the fall in demand can be eliminated by a rise in W_0 and p_0. Thus the FF schedule shifts to the right to $F'F'$. As is evident the horizontal shift of the FF schedule is proportional to $(1-\gamma)$; if $\gamma=1$ the schedule and the initial equilibrium remain intact (the Ricardian equivalence case). The new equilibrium obtains at point B with a *higher rate of interest, a higher relative price of nontradable goods* p_0, *and a higher level of domestic wealth and consumption. The higher rate of interest lowers foreign wealth and consumption and reduces the foreign relative price of nontradable goods.* Thus, on the basis of the correlations between domestic and foreign private-sector spending and between domestic and foreign real exchange rates, *the international transmission of the budget deficit is negative.*

As an interpretation, we note that since the budget deficit transfers income from future generations (whose propensity to consume present goods is zero) to the current generation (whose propensity to spend on present goods is positive), it creates an excess demand for present tradable goods resulting in a rise in their *intertemporal* relative price (the rate of interest). Likewise, it creates an excess demand for domestic nontradable goods and an excess supply of foreign nontradable goods and changes the *temporal* relative prices (the real exchange rates). Generally speaking, this pattern of consumption, real interest rates, real exchange rates, and the underlying fiscal positions is roughly in accord with the selected facts reported in Section I (for a related analysis, see William Branson, 1985).

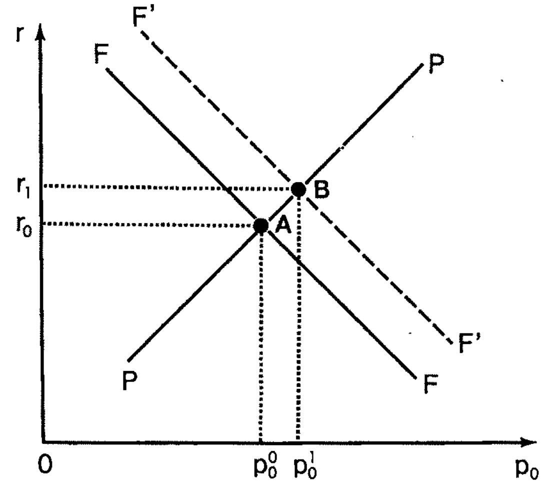


FIGURE 1. BUDGET DEFICITS, THE REAL EXCHANGE RATE, AND THE REAL RATE OF INTEREST

A key characteristic of the conceptual framework underlying the model is that it is forward looking. Hence, the timing of policy actions plays a critical role. To illustrate this point, we apply a simplified version of the model in which the economy produces only tradable goods, to an analysis of transitory and permanent balanced-budget changes in government spending. In that case, equations (1) and (2) are modified in an obvious manner,[2] and the equilibrium is illustrated

[2] In the absence of nontradable goods we eliminate the subscript T, we set $\beta=\beta^*=0$, $I[\cdot]=Y-T$, $I^*[\cdot]$

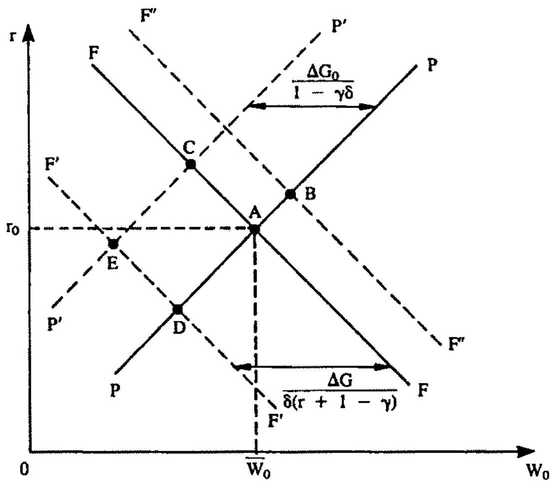


FIGURE 2. CURRENT AND FUTURE GOVERNMENT SPENDING, WEALTH, AND THE REAL RATE OF INTEREST

by point A in Figure 2. The positively sloped PP schedule shows combinations of r and W_0 that clear the market for present goods. The negatively sloped FF schedule describes combinations of r and W_0 that clear the market for future goods. In this simplified version of the model a current tax cut shifts the FF schedule to the right (to $F''F''$) and, as before, in the new equilibrium (point B) the rate of interest and domestic wealth are higher. A transitory rise in *current government spending* by ΔG_0 creates an excess demand for present goods (since the private-sector propensity to spend on current goods is $1-\gamma\delta$) and raises the rate of interest.

Diagrammatically, the PP schedule shifts to the left by $\Delta G_0/(1-\gamma\delta)$ to $P'P'$ and the new equilibrium obtains at point C. Analogously, a balanced budget rise in *future government spending* by ΔG creates an excess demand for future goods, shifts the FF schedule to the left by $\Delta G/[\delta(r+1-\gamma)]$ to $F'F'$, and lowers the rate of interest. The new equilibrium obtains at point D. In the former case both domestic and foreign wealth fall and *the transmission is positive*; in the latter case domestic wealth falls, foreign wealth rises and *the transmission is negative*. A *permanent* balanced-budget rise in government spending raises demand for both present and future goods and shifts both schedules (with $\Delta G_0 = \Delta G$). The impact on the rate of interest depends on the *relative* excess demands in both markets. If the home country was a net saver (i.e., if $\delta > 1/(1+r)$, or equivalently if $\delta > \delta^*$), the permanent rise in government spending raises the relative demand for present goods and the rate of interest rises; in that case foreign wealth falls. The opposite, illustrated by point E in Figure 2, holds if $\delta < \delta^*$. These results suggest that the (apparently unstable) relations between government spending, real rates of interest, and the international transmission can be explained in part in terms of different expectations concerning future spending.

In order to analyze the effects of *future* budget deficits, we modify the specification of the time aggregation of the model and divide the horizon into the present, the near future, and the distant future. It can be shown (see our article, 1986a) that, analogously to the effects of current deficits, a tax cut in the near future (followed by a corresponding tax rise in the distant future) creates an excess demand for goods in the near future, and, raises the *future* rate of interest, domestic wealth, and spending while lowering foreign wealth and spending. Thus, *the transmission of future budget deficits is negative*. Their impact on the *current* short-term rate of interest depends on the saving propensities; if $\delta < \delta^*$ the current short-term interest rate rises, and vice versa.

In interpreting this result, we note that in the present period no government action takes place and changes in the current rate of interest result only from changes in world savings. At the prevailing short-term interest rate, foreign wealth falls because of the rise in the future rate of interest while the rise in domestic wealth consequent on the future budget deficit is mitigated by the rise in the future rate of interest. These changes in wealth lower the foreign demand for current goods and raise the domestic demand for these goods. *World* demand for current goods rises or falls depending on the difference

$= Y^* - T^*$, $Y^*_{N0} = Y^*_N = 0$, and $W^*_0[\cdot] = Y^*_0 - T^*_0 + (Y^* - T^*)\gamma/\rho + B^*_{g0} + B_0$. The resulting model is analyzed in detail in our article (1986a).

between the two spending propensities. The unambiguous fall in foreign wealth indicates that even though the current short-term rate of interest may fall, the future budget deficit must raise the overall "appropriate average" of short- and long-term rates of interest. In this context, recall from Section I that, in recent years, changes of long-term real rates of interest exceeded those of short-term rates. In addition to being induced by other factors, this fact can result in part from expectations of future large U.S. budget deficit.

In summary, the model offers predictions about the intercountry correlations among private-sector spending as well as about the links between fiscal policies, real exchange rates, and world real interest rates. It was shown that budget deficits arising from current or future *tax cuts* result in *negative* intercountry correlations among private consumption. On the other hand, the correlations implied by changes in government *spending* depend on the timing of these changes and on the current-account positions of the various countries. It was also shown that a budget deficit arising from a tax cut *raises* real interest rates linking the period of the tax cut and the future. The effect on the current short-term rate of interest of either a future budget deficit or of permanent changes in government spending depend on the current-account positions of the various economies. Finally, the current short-term real rate of interest rises in response to a current transitory rise in government spending and falls in response to a future transitory rise in government spending.

Finally, it is important to emphasize that by focusing on fiscal policies and by excluding monetary considerations, the analytical framework is limited. As a result, although the analysis accounts for some of the facts outlined in Section I, it does not provide an explanation for the timing of the initial rise in real rates of interest (in the late 1970's and the beginning of the 1980's) and the timing of the decline in real rates since mid-1984. The likely explanations for these facts can be given in terms of U.S. monetary policy. Therefore, a useful extension would include monetary considerations.

REFERENCES

Blanchard, Olivier J., "Debt, Deficits and Finite Horizons," *Journal of Political Economy*, April 1985, *93*, 223–47.

Branson, William H., "Causes of Appreciation and Volatility of the Dollar," NBER Working Paper, No. 1777, 1985.

Frenkel, Jacob A. and Razin, Assaf, (1986a) "Fiscal Policies in the World Economy," *Journal of Political Economy*, forthcoming June 1986.

______ and ______, (1986b) "Real Exchange Rates, Interest Rates and Fiscal Policies," *Economic Studies Quarterly*, forthcoming August 1986.

International Monetary Fund, *World Economic Outlook*, Washington, October 1985.

The Uneasy Case for Greater Exchange Rate Coordination

By JEFFREY SACHS*

The September 1985 agreement of the finance ministers of the G-5 (United States, United Kingdom, Germany, France, Japan) to cooperate in producing an "orderly appreciation of the main non-dollar currencies against the dollar" appears to have ushered in a new phase of international monetary coordination. The agreement had its desired effect in the short term, by bringing about a 10 percent trade-weighted depreciation of the dollar in the 2 months following the accord.[1] However, while the specific announcement by the G-5 ministers was no doubt successful, and while most observers had called for a decline in the dollar in the face of growing protectionist pressures in the United States, the longer-term aspects of the G-5 agreement are problematic to say the least. The ministers agreed on a target, but said almost nothing about how policies should be implemented in the future in pursuit of that target.

Choosing ends before means is always dangerous, and clearly so in matters of exchange rate management. Economic history shows that the general form of an exchange rate system can be less important than the detailed policies that are implemented in conjunction with the system. Advocates of fixed rates, for example, should reflect on the disasterous character of the Smithsonian Agreement of December 1971, hailed at the time by President Nixon as "the most significant monetary agreement in the history of the world." The preceding Bretton Woods arrangements had worked well until U.S. monetary policy became overly expansionary. The Smithsonian Agreement tried to sustain the success of Bretton Woods without solving the fundamental problem of U.S. policy. In the end, the arrangement collapsed after 14 months, and only after Germany and Japan had imported an inflationary dose of U.S. monetary expansion. Rather than providing "discipline," as fixed exchange rate systems are often alleged to do, the Smithsonian Agreement played a key role in ushering in the high inflation of the 1970's.

In the current circumstances, U.S. fiscal policy rather than monetary policy is the predominant factor putting stress on exchange rates. The G-5 accord, like the Smithsonian Agreement, has so far settled on form, rather than the content of realigning fiscal policies. As in 1971-73, there are signs that the agreement could do more to "export" our fiscal deficits to Germany and Japan, which are presently trying to reduce their own deficits, than to set the major economies on a path of more stable policies. In a nutshell, the convergence of policies implied in "managed" exchange rate arrangements can as easily be convergence around a bad set of policies as around a good set of policies.

This note reviews the theoretical and historical case for managed exchange rates (a more detailed treatment of many of these issues can be found in my 1986 paper). After reviewing the general case for a move towards greater management, the specific circumstances surrounding the G-5 accord are discussed.

I. Arguments for a More Managed Exchange Rate System

The theoretical arguments for and against managed exchange rates have been richly debated in the economics literature for decades. While the theoretical sophistication and historical experience have developed extensively over time, the core of the arguments have remained approximately the same. First, advocates of managed rates argue that the foreign exchange markets are subject

*Harvard University, Cambridge, MA 02138, and NBER.

[1]The data are from the Bank of England, as reported in *The Economist*, financial indicators, September 27, 1985, and January 3, 1986.

to speculative bubbles and runs that make a purely floating rate a poor guide to resource allocation. Managed rates are alleged to be capable of keeping exchange rates closer to the fundamentals. Second, managed exchange rate systems are seen by advocates as providing necessary "rules of the game" to limit the discretion of national policy authorities. Without rules of the game, it is argued, national policymakers are likely to pursue policies that are harmful to other countries, and that may even be harmful to their own economy. Third, advocates of managed rates regard reduced volatility of exchange rates an important end in itself. A managed exchange rate system moves the world closer to the norm of an international money, with efficiency gains that are judged to be analogous to a unified currency on the national level.

Each of these three arguments for managed rates has been fiercely debated in the literature. The first argument, about the inefficiencies of the foreign exchange market itself, have long been at the center of the debate. Ragner Nurkse (1944) made the assertion, widely adopted, that floating rates in the 1920's had been undone by destabilizing speculation. Milton Friedman (1953) in a famous retort, argued that destabilizing speculators would necessarily lose money, and thereby be weeded out of the market. More recent theoretical studies have stressed that even with rational market participants, exchange markets can be subjected to whims and fads that are not grounded in the fundamentals, and that indeed can impose social costs (see Olivier Blanchard, 1979; Maurice Obstfeld, 1986; Robert Flood and Peter Garber, 1984, as examples). "Rational bubbles," "sunspot equilibria," and "self-fulfilling runs" are all cases of destabilizing speculation by rational investors.

The second argument for managed rates seeks to limit the autonomy of national policymaking by imposing international rules of the game. There are really two distinct arguments in favor of international rules. First, it is feared that if national policies are chosen in an unconstrained manner, policymakers will select policies that "beggar thy neighbor." As in the prisoners' dilemma, the result will be a loss of economic welfare for all countries concerned. An example of such an inefficient policy outcome is presented by Warwick McKibbin and me (1986), in which the major OECD regions are simultaneously confronted with an inflationary shock. Each of the regions attempts (in vain) to appreciate its currency relative to the others, in order to export some of the inflationary shock. The end result is an OECD policy mix with excessive monetary contraction cum fiscal expansion. The second motivation for rules of the game is the fear that national authorities cannot be trusted to pursue even their own national advantage, since they are prone for a variety of reasons to choose overly inflationary policies. The political business cycle models, and the positive theory of inflation offered by Robert Barro and David Gordon (1983), both illustrate the inflationary bias. International rules of the game are seen as useful in restraining such behavior.

Unfortunately, the benefits from tighter rules of the game cannot be guaranteed, for several reasons. Theoretical models by Koichi Hamada (1976), and the historical analysis of the interwar period by Barry Eichengreen (1985), show that even a fixed exchange rate system does not necessarily eliminate the prisoners' dilemma aspects of national policymaking. Under floating exchange rates, policymakers might try to manipulate the exchange rate to their national advantage; under fixed rates, they might instead try to manipulate the rate of reserve accumulation or some other policy variable. Nor do tighter rules of the game necessarily overcome the inflationary bias of national policymakers. If national policymakers lean towards overly inflationary policies, global policy coordination by these same policymakers might make the inflationary bias even worse! Kenneth Rogoff (1985), for example, has presented an illustration in which national policymakers have an overinflationary bias, but one that is held in check by the fear of each policymaker that the national currency will depreciate if the policies are more inflationary than abroad. When these policymakers undertake jointly to manage their exchange rates, they are thereby freed to have a com-

mon inflationary expansion, without fear of depreciation against each other!

The third argument for managed rates holds that exchange rate stability per se is a public good, as is a unified currency within a single country. Charles Kindleberger, a leading advocate of this position, has argued:

> The main case against flexible exchange rates is that they break up the world market...Imagine trying to conduct interstate trade in the USA if there were fifty different state monies, no one of which was dominant. This is akin to barter, the inefficiency of which is explained time and again by textbooks. [1981, p. 174]

This argument must be balanced, as in the optimal currency area literature, by the argument that movements in nominal exchange rates can provide a very efficient way to respond to disturbances that require relative price changes across countries.

II. The G–5 Agreement in View of the Theoretical Arguments

Decades of debate over exchange rate arrangements have not produced a clear consensus on the "optimal" system. I have noted that sound theoretical arguments lie on both sides of the debate over flexible vs. managed rates. The result is that the case for one system over another is likely to be historically determined, and to depend on: the depth and efficiency of the foreign exchange markets at a period of time; the quality of political leadership and accountability in the major economies; and on the nature of the economic disturbances and dislocations which are confronting the economies. In this spirit, it is important to ask whether current conditions in fact point to the need and benefit of greater exchange rate management.

The period of floating has been dominated by large and sustained deviations of exchange rates from levels consistent with purchasing power parity (*PPP*), with the real appreciation of the dollar since the end of 1980 being the most remarkable and important case. At the same time, the period has witnessed a marked divergence in macroeconomic policies in the major economies, the most notable of course being the sustained U.S. fiscal expansion since 1980 in contrast with the sustained fiscal contraction in Europe and Japan. Do the extreme movements in the dollar provide a prima facie case for greater exchange rate management? According to the theoretical arguments, we should first ask whether the extreme movements of the dollar should be regarded as a "bubble," rational or otherwise, or should be traced instead to fundamentals, particularly to divergent fiscal policies. The evidence suggests that most, if not all, of the appreciation of the dollar has been tied to fundamentals, or has at least been consistent with the developments in other asset markets, particularly the bond market (see my 1985 paper and Jeffrey Frankel, 1985 for this view, and Paul Krugman, 1986, for a partial dissent). Bubbles probably help to explain some short-term volatility of exchange rates (as in T. Wing Woo, 1984), rather than persistent deviations from *PPP* over a period of several years.

If exchange rates are tied to fundamentals, then the extent of the dollar's real appreciation should depend on the long-term real interest rate differential between the United States and the rest of the OECD. A one-percentage point real differential in favor of the United States on 10-year bonds should be accompanied, approximately, by a 10 percentage point real appreciation of the dollar relative to its long-run level (see my 1985 paper, p. 148). Econometric and less formal evidence in my paper and Frankel indeed support the interpretation that the movements of the dollar were closely tied to interest rate differentials. Furthermore, simulation models have shown that the divergence in fiscal policies since 1980 can plausibly account for the movement in long-term interest rate differentials, and therefore for the movement in the dollar. Using a model that McKibbin and I developed, I found that a 4 percent of *GNP* U.S. fiscal expansion, combined with tight U.S. monetary policy, and a 2 percent of *GNP* fiscal contraction in the rest of the OECD, combined with unchanged monetary policy,

would result in a real appreciation of the dollar of about 39 percent (see my 1985 paper, p. 175).

These findings would seem to be contradicted by the results of the G–5 agreement itself, which appeared rather effortlessly to bring down the value of the dollar by about 10 percentage points. Many have interpreted the initial success of the G–5 agreement rate management to prove that indeed an exchange rate bubble was burst, without the need for policy intervention. However, this view is illusory. The G–5 announcement itself provided important signals about future policies in the United States, Germany, and Japan. At the very least, the United States committed itself to tie its monetary policy in part to an exchange rate target, rather than to a strict money growth rule. This commitment was plausible in view of the low inflation and weak growth of the U.S. economy since mid-1994. Japan, on the other hand, committed itself to tighter monetary policy. The result has been some convergence in short- and long-term interest rates in the major economies in the 2 months following the agreement. The U.S. short rates have declined by about 50 basis points, while Japanese and German short rates have risen by 115 and 15 basis points, respectively. On the long end, U.S. government bond yields have declined by about 100 basis points, while Japanese and German government bond yields have risen by 10 and 35 basis points. These long-term interest rate movements can account roughly for a 11 percent appreciation of the yen vis-à-vis the dollar, and a 13 percent rise in the deutsche mark, according to the rule of thumb enunciated earlier.[2]

Will the accord help move the system towards greater discipline, as advocates of international exchange rate management suggest? Here, the evidence is of course fragmentary, but not entirely encouraging. The passage of the Gramm-Rudman deficit reduction legislation signals some movement on U.S. deficits, and is probably one factor behind the fall in long-term U.S. interest rates. But in addition, following the accord, the United States has been prodding Germany and Japan to *increase* their interest rates, through tight money and larger budget deficits. This aspect of policy coordination would be self-destructive as well as shortsighted. Neither country should be raising interest rates on the basis of internal conditions. With respect to budgets, Germany and Japan have undertaken 4 years of politically painful adjustments on their fiscal balance in order to reduce general government deficits from around 4 percent of *GDP* to around 1.5 percent of *GDP* in 1985. (Data are from the *OECD Economic Outlook*, July 1985.) From 1970 to 1985, a sustained period of deficits in these two countries raised the public debt/*GNP* ratio from about 20 to 40 percent in Germany, and from about 5 to 40 percent in Japan. For the United States to be urging a reversal of the fiscal discipline at a time when it has itself recognized the urgency of fiscal restraint is irresponsible.

A reversal of fiscal restraint or a tightening of monetary policy in the rest of the OECD is neither necessary from a cyclical point of view, nor conducive to longer-term cooperation of exchange rates. It is sometimes suggested that a fiscal contraction in the United States would require a fiscal expansion abroad in order to keep OECD output growth unchanged, but this ignores the fact that a fiscal contraction in the United States can instead be matched by a monetary expansion in the United States and abroad of sufficient magnitude to maintain an unchanged output path. Using McKibbin's and my simulation model, a 3-year phased reduction of the U.S. deficit by 1 percent of *GNP* each year, combined with offsetting monetary expansions in the various OECD regions, produces a package which: is neutral with respect to output growth; causes the dollar to depreciate by 14 percent over 3 years; improves the U.S. current account by 1.6 percent of *GDP*; and reduces nominal short-term interest rates in Japan and the rest of the OECD by

[2] The data are from *The Economist* financial indicators (see fn. 1). The short rates are 3-month money market rates, and the long rates are for government bonds. The calculation in the text simply multiplies the change in the long-term interest rate differential by 10 to get the "predicted" exchange rate effect.

about 3 percentage points. The key result is that there is no obvious reason to supplement the package with fiscal expansions abroad.

There is little guarantee that more management would result in greater discipline, on average, particularly if it pulls the "responsible" countries in the direction of greater budget deficits. This possibility is no doubt behind the manifest lack of enthusiasm shown by Germany and Japan for reform of the exchange rate system. While it is true that we can conceive of international agreements that would in principle lead to better behavior, we should take Jacob Frenkel's warning (1984, p. 122) to heart not to compare a "best" managed system with a poorly run flexible rate system. Prudence requires us to compare worst cases with worst cases.

The strongest long-term case for greater exchange rate management probably lies with the third argument I outlined earlier: the importance of an international money. Robert Mundell (1985) and Kindleberger both make the point that the core of stability of the world system in the past 15 years has not been the efficiency of a multicurrency world, but rather the dominance of the dollar in world trade. As Mundell states, "Underneath the stormy float is the undercurrent of a vast dollar area.... The dollar has provided the backbone of the international monetary system since the breakdown of the gold exchange standard in 1968 and 1971" (p. 2). The growing crisis, however, is that with the U.S. economy shrinking as a share of world income, and increasingly vulnerable to external shocks, the dollar may be unable to play its assigned role in future years. Confidence in the dollar as the world currency will also be deeply undermined by the transformation of the United States into the world's largest debtor economy in the next few years. For these reasons, stability in trade will require a much more conscious effort at reduced volatility among the dollar, and those currencies, such as the yen and the deutsche mark, which will compete with the dollar as world currencies.

Since exchange rate management per se will probably do little to result in better national policies, it is important to set new rules of the game ahead of making commitments to set exchange rate targets. While many theoretical models have suggested how such rules might be designed, there has so far been little empirical work on the operational properties of alternative rules for exchange rate management (see McKibbin and myself for a rough start). This daunting task should have a high priority in future research in international economics.

REFERENCES

Barro, Robert and Gordon, David, "Rules, Discretion, and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, *12*, 101–21.

Blanchard, Olivier, "Speculative Bubbles, Crashes and Rational Expectations," *Economics Letters*, 1979, *3*, 386–89.

Eichengreen, Barry, "International Policy Coordination in Historical Perspective: A View from the Interwar Years," in W. Buiter and R. Marston, eds., *International Economic Policy Coordination*, Cambridge: Cambridge University Press, 1985.

Flood, Robert and Garber, Peter, "Gold Monetization and Gold Discipline," *Journal of Political Economy*, February 1984, *92*, 90–107.

Frankel, Jeffrey, "The Dazzling Dollar," *Brookings Papers on Economic Activity*, 1:1985, 199–218.

Frenkel, Jacob, Discussion of "Exchange Rate Arrangements in the Eighties (by Robert Roosa)," in *The International Monetary System*, Federal Reserve Bank of Boston Conference Series, No. 28, 1984.

______, *International Aspects of Fiscal Policies*, National Bureau of Economic Research, forthcoming 1986.

Friedman, Milton, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.

Hamada, Koichi, "A Strategic Analysis of Monetary Interdependence," *Journal of Political Economy*, August 1976, *84*, 677–700.

Kindleberger, Charles, *International Money*, London: Allen & Unwin, 1981.

Krugman, Paul, "Is the Strong Dollar Sustainable?," 1985 conference volume, Federal Reserve Bank of Kansas City, forthcoming 1986.

McKibbin, Warwick and Sachs, Jeffrey, "Coordination of Monetary and Fiscal Policies in the OECD," in J. Frenkel, ed. *International Aspects of Fiscal Policies*, NBER, forthcoming 1986.

Mundell, Robert, "Proposals for the Future International Economic System," presented at the Congressional Summit on Exchange Rates and the Dollar, Washington, D.C., November 1985.

Nurkse, Ragner, *International Currency Experience, Lessons of the Interwar Period*, Princeton: League of Nations, 1944.

Obstfeld, Maurice, "Rational and Self-Fulfilling Balance-of-Payments Crises," *American Economic Review*, March 1986, *76*, 72–81.

Rogoff, Kenneth, "Can International Monetary Policy Cooperation Be Counterproductive," *Journal of International Economics*, February 1985, *18*, 199–217.

Sachs, Jeffrey, "The Policy Mix and the Dollar: 1985," *Brookings Papers on Economic Activity*, 1:1985, 117–97.

______, "Is There a Case for More Managed Exchange Rates?" 1985 conference volume, Federal Reserve Bank of Kansas City, forthcoming 1986.

Woo, Wing T., "Speculative Bubbles in the Foreign Exchange Markets," *Brookings Discussion Papers in International Economics*, March 1984.

U.S. Budget Deficits and the European Economies: Resolving the Political Economy Puzzle

By MARTIN FELDSTEIN*

The experience of the past 5 years has dramatically underscored the international consequences of U.S. fiscal policy. The massive current and projected budget deficits were the primary cause of the sharp 60+ percent rise in the real trade-weighted value of the dollar between 1980 and the end of 1984. The dollar's rise was in turn the major reason that the U.S. current account shifted from a surplus in 1981 to an unprecedented deficit that currently exceeds 3 percent of *GNP*.[1]

During the past 4 years, the exploding trade deficit has generated calls in the United States to adopt protectionist policies and to abandon our basic system of floating exchange rates. At the same time, European policy officials have been unanimous in calling upon the United States to reduce its budget deficit and to cooperate in intervention to lower the dollar's value.

The anguished cries from American exporters and from American firms and employees that compete with imports from abroad are easy to understand. But why have European officials been critical of our fiscal policy and eager to see the dollar decline? Hasn't the U.S. trade deficit meant a substantial improvement in the European trade balances? And shouldn't that have been the mechanism by which the U.S. "locomotive" (to use a currently unpopular phrase) pulled the European economies out of their recession? The European complaints about U.S. fiscal policy and about the resulting strong dollar are a political economy puzzle that deserves to be addressed.

There is no doubt that the European trade balances have improved. For the European Economic Community (EEC) as a whole, the trade balance with the rest of the world improved by more than $50 billion between 1980 and 1984. Exports rose sharply. French and German exports both increased about 17 percent between 1980 and 1984 and have continued to expand rapidly in 1985. Nevertheless, during the past 4 years, growth in the European community languished at an average annual rate of less than 1 percent and the unemployment rate has doubled.

It is of course possible that the rise in European unemployment and the slow growth in Europe would have been even worse without the strong U.S. dollar and our resulting trade deficit. If so, the vociferous complaints of European officials are simply the normal behavior of politicians seeking to shift the blame for domestic policy failures to a convenient scapegoat. But I think that there is a more serious and legitimate reason why the Europeans have been critical of our fiscal policy.

In considering the impact of recent U.S. fiscal policy on the European economies, it is important to bear three things in mind. First, budget policy is only one aspect of the change in U.S. economic policy that has occurred since 1980. The increased budget deficits were accompanied by a monetary policy that caused U.S. inflation to decline from 9 percent in 1980 and 1981 to 4 percent in each year since 1982. Thus a sound monetary policy and the expectation that such a policy would persist prevented the massive budget

*Professor of Economics, Harvard University, Cambridge, MA 02138, and president of the National Bureau of Economic Research. I am grateful to Rudiger Dornbusch, Jeffrey Frankel, Paul Krugman and Jeffrey Sachs for conversations on this subject over the past several years in both Washington and Cambridge. A more complete discussion of the impact of U.S. deficits on German unemployment will appear in my paper with Philippe Bacchetta (1986). Earlier comments along these lines were published in my 1985 paper.

[1]For a discussion of the links between the budget deficit, the dollar and the trade deficit, see ch. 3 of the 1983 and 1984 editions of the *Economic Report of the President*, and my paper (1986b).

deficits from leading to concurrent inflation or to a widespread expectation of future inflation.[2] In addition, business tax rules were changed in 1981 in a way that reinforced the effect of the budget deficits on real interest rates.[3]

Second, U.S. policies affect the European economies not only directly through trade and financial markets, but also by inducing changes in the policies of European governments. These induced policy changes have been particularly important in the past few years and go a long way toward explaining why European officials believe that their countries have been hurt by U.S. fiscal policies despite the favorable effects on European exports.

Third, although it is possible to generalize considerably about the impact of U.S. policy on the European economies, it is also important to recognize that the situation differs among the major countries. I will focus my comments on the situation in Germany, not only because Germany is the largest of the European nations, with nearly 30 percent of the EEC's total *GDP*, but because it is also the dominant nation in the European Monetary System and the primary trading partner of the other European nations. The impact of the U.S. deficits on the German economy and particularly on German economic policy therefore had profound effects elsewhere in Europe.

I. The Direct Impact of U.S. Fiscal Policy

The strong U.S. economic recovery and the surge in the U.S. dollar undoubtedly contributed to a significant rise in European exports to the United States. Since 1980, U.S. imports from Western Europe increased more than 50 percent while the value of our exports to those countries has actually declined. But the importance of this to Europe should not be exaggerated. The United States represents less than 10 percent of the imports and exports of the EEC countries; even when intra-EEC trade is eliminated, the United States is only one-sixth of EEC trade. In fact, the total volume of exports of the EEC rose only 14 percent between 1980 and 1984, about one-third less than the increase in the two previous 4-year periods.

The sharp rise in the dollar relative to the European currencies also had two unfavorable direct effects. It caused a decline in the terms of trade with the United States (and with other countries whose currencies followed the dollar more closely than the ECU) and it put upward pressure on domestic inflation in Europe. The German mark fell 45 percent in real terms relative to the dollar between 1980 and the end of 1984, implying that German exports to the United States bought 45 percent less in U.S. goods in 1984 than they did in 1980. Since U.S. trade represents 8 percent of Germany's *GDP*, this corresponds to a real income fall of about 3 percent.

The impact of the dollar's rise on European inflation is more difficult to judge because of the induced policy response to which I will turn in a moment. But a simple mechanical pass-through of the increased import costs associated with the 70 percent nominal rise in the dollar-deutsche mark ratio between 1980 and the end of 1984 would imply a 6 percent increase in the German price level or about one-third of the actual price rise over that 4-year period. This cost pass-through calculation is very crude. It understates the inflationary effect to the extent that it ignores imports priced in dollars that do not come from the United States, and because it ignores the impact of rising import prices on domestic prices and wages.

The increase in U.S. real long-term interest rates raised not only the value of the dollar, but also the level of real long-term interest rates in Europe. This increase in interest rates reflected in part the policy response of the Bundesbank and other central banks that I will discuss in a moment. But even with no

[2] This gives a bit too much credit to monetary policy for the decline in U.S. inflation. The interaction of monetary and budget policy, by raising the exchange value of the dollar, was an important cause of reduced inflation. See Jeffrey Sachs (1985).

[3] See my papers (1976, 1980) for an early discussion of the likely impact of such tax reform on real interest rates, and my paper (1986a) for evidence that the enlarged budget deficits had a much bigger impact on interest rates than the change in tax rules.

change in European economic policy, the behavior of private investors would have caused the real long-term interest rates on the deutsch mark and other long-term bonds to move in the same direction as the real interest rate on dollar bonds. The rise in the real long-term interest rate in Germany reduced spending on construction and on investment in plant and equipment, industries in which the decline in activity and increase in unemployment have been particularly severe.

This impact of U.S. long-term rates on European investment can also be expressed in terms of the shift in the capital flow with the rest of the world. Germany, which was importing capital to finance a current account deficit in 1980, had a current account surplus in 1984 equal to more than 5 percent of gross fixed investment. For 1985, the current account surplus and associated capital outflow is expected to reach more than 10 percent of gross fixed investment. Europeans complain that this not only represents a fall in activity in the investment sector but also, by keeping the capital stock smaller than it would otherwise have been, retards the growth of productivity and therefore exacerbates the classical unemployment problem caused by wage demands in excess of full-employment productivity.

It is difficult to be confident about the net impact on employment and output of the several direct effects of the U.S. fiscal deficits. The positive impact of the expanded exports may, but need not, exceed the adverse effects of the higher real interest rates and capital outflow. But these direct effects are only part of the total impact of U.S. fiscal policy. The large and persistent U.S. budget deficits also caused the European countries to alter their monetary and budget policies in a contractionary way. To understand the impact of U.S. deficits on European economic activity, it is important to examine the nature of these induced changes in economic policy.

II. The Induced Policy Response

Inflation was a principal problem on the minds of policy officials everywhere in Europe as the 1980's began. The rise in the inflation rate after the 1973 oil shock had just been slowly but painfully reversed when the 1979 oil shock restarted the inflation process. Inflation in the EEC as a whole rose from 6.9 percent for 1978 to 12.4 percent for 1980. The Germans saw inflation rise from 2.6 percent for 1978 to 5.3 percent for 1980. The German government was determined to reverse this inflation, even if it meant temporarily slower growth and increased unemployment.

The sharp rise in the dollar that began in 1980 brought with it an additional inflationary impulse: a 20 percent decline in the mark between 1980 and 1981. To limit the future decline in the mark, the Bundesbank slowed the growth of money and raised short-term interest rates. The German money supply actually declined between 1980 and 1981. The mark nevertheless continued to fall. To offset the resulting inflationary pressure, the Bundesbank appears to have continued to manage monetary policy in a way that maintained greater slack in the German economy than they would otherwise have preferred. Money supply growth from 1980 to 1983 was approximately half of what it had been in the three preceding 3-year periods and real *GNP* grew at an average annual rate of less than 1 percent. This contraction of German monetary policy in reaction to the rise of the dollar caused by U.S. budget deficits was an important source of continued high unemployment in Germany and elsewhere in Europe.

All of this can be restated more formally as follows: the decline in the mark that resulted from the rising U.S. budget deficits was a negative supply shock to the German economy that shifted its short-run Phillips curve to the right. The Bundesbank pursued a tight money policy aimed at limiting this shift and, to the extent that the shift could not be prevented, at moving along the Phillips curve to achieve lower inflation than would otherwise occur but at the cost of higher unemployment.

The enlarged U.S. budget deficits also appear to have caused German fiscal policy to become more contractionary. This occurred for several reasons. When the combination of high German interest rates and a slack economy automatically enlarged the budget

deficit, the German government sought to counter this by cutting spending and allowing tax receipts to rise as a share of income. It was also argued that shrinking the German budget deficit would reduce inflationary pressure in Germany, would lower real interest rates (including those paid by the government), and would free up funds to offset the capital outflow.

These remarks are not meant as either praise or criticism of German monetary and fiscal policy since 1980. I want only to indicate how the enlarged U.S. budget deficits, by raising real U.S. interest rates and thereby reducing European currency values, induced the German government to adopt contractionary monetary and fiscal policies. Because of Germany's dominant position in the European Monetary System, the other countries of Europe had to follow Germany's tight money policy.[4] Moreover, because half of European trade is within the EEC, and Germany is the largest EEC market for each of the other member countries, the slowdown in German economic activity was immediately transferred throughout Europe.

III. The Falling Dollar

The fall of the dollar that began in March 1985 opened new opportunities for Germany and the other European countries. The dollar fell from 3.4 marks at its peak to 2.8 marks in September (before the G–5 announcement on coordinated intervention) and to less than 2.6 marks by the beginning of December. Parallel shifts occurred in the values of the other EMS currencies.

The preceding analysis suggests that Germany and the other EEC countries would respond to this dollar decline by easing monetary policy. That is exactly what occurred. Although German money market rates were at essentially the same level in March 1985 as they had been at the start of 1984 (despite a 100 basis-point decline in comparable U.S. rates and a significant decline in German inflation), during the 6 months after the dollar began to decline, German interest rates came down 150 basis points, widening the gap with U.S. rates. The response was similar in other EMS countries: money market rates have declined 150 basis points in France and Italy. Longer-term rates have also generally declined.

The G–5 decision to try to bring the dollar down by coordinated intervention has started yet another new phase. Although the Japanese have strengthened the yen more than 15 percent by a sharp tightening of monetary policy and by increasing the likelihood of fiscal expansion, the Germans and other key European governments have only engaged in limited intervention without any noticeable shift in monetary or fiscal policies. The more modest strengthening of their currencies relative to the dollar since September has reflected a continuation of the trend started earlier in the year, accelerated by the market's fear of possible intervention and perhaps by the growing recognition that the dollar remains too high to decline only at a rate equal to the international interest differential (Paul Krugman, 1985).

The strengthening of the European currencies relative to the dollar replaces imported inflation with imported disinflation. In this environment, the European governments can more easily pursue expansionary monetary policies. This opportunity will be reinforced if prospective U.S. budget deficits and long-term interest rates continue to decline. Although such an easing of European monetary policy cannot be a substitute for dealing with the more fundamental problems that prevent a return to full employment in Europe, it can help to reduce the current exceptionally high levels of European unemployment.

[4] This induced shift of national monetary policies in response to U.S. fiscal deficits is the explanation of the rise in worldwide interest rates that puzzled Olivier Blanchard and Lawrence Summers (1984). Malcom Knight and Paul Masson (1985) indicate that a substantial rise in worldwide interest rates can also be ascribed to the net fiscal expansion that occurred in the world economy as U.S. deficits increased by more than European deficits declined. The fall in the OPEC surplus reinforced this net decline in worldwide savings.

REFERENCES

Blanchard, Olivier and Summers, Lawrence, "Perspectives on High World Real Interest Rates," *Brookings Papers on Economic Ac-*

tivity, 2:1984, 273–324.

Feldstein, Martin, "Inflation, Income Taxes and the Rate of Interest: A Theoretical Analysis," *American Economic Review*, December 1976, *66*, 809–20.

______, "Tax Rules and the Mismanagement of Monetary Policy," *American Economic Review Proceedings*, May 1980, *70*, 182–86.

______, "American Economic Policy and the World Economy," *Foreign Affairs*, Summer 1985, *63*, 995–1008.

______, (1986a) "Budget Deficits, Tax Rules and Real Interest Rates," NBER working paper, forthcoming 1986.

______, (1986b) "The Budget Deficit and the Dollar," in NBER *Macroeconomics Annual 1986*, forthcoming 1986.

______ **and Bacchetta, Philippe,** "U.S. Budget Deficits and European Unemployment: The German Experience," NBER working paper, forthcoming 1986.

Knight, Malcolm and Masson, Paul R., "Fiscal Policies, Net Savings and Real Exchange Rates: The United States, Japan and the Federal Republic of Germany," paper presented to a NBER conference on *International Aspects of Fiscal Policies*, December 1985.

Krugman, Paul, "Is the Strong Dollar Sustainable?," Working Paper No. 1644, NBER, 1985.

Sachs, Jeffrey, "The Dollar and the Policy Mix: 1985," *Brookings Papers on Economic Activity*, 1:1985, 117–97.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, 1983; 1984.

The Black Underclass Concept: Self-Help vs. Government Intervention

By EMMETT D. CARSON*

The concept of an American underclass has become the subject of increasing discussion. Although the term does not enjoy a precise definition, most researchers acknowledge that the underclass condition involves more than being cash poor. Members of the underclass are also believed to have attitudinal and behavioral deficiencies. Other terms that have been used to describe this group over the years include: "dangerous classes," "other Americans," "culture of poverty," and "lower classes" (see Ken Auletta, 1981). Within this area of inquiry, several researchers have chosen to study what has been termed the "black underclass." The purpose of this paper is twofold. First, the appropriateness of the data that are used to characterize members of the black underclass will be examined. Second, the argument that perhaps the most viable solution to the problems of the black underclass are community self-help programs, rather than government programs, will be discussed.

I. Evidence of a Black Underclass?

There have been only a handful of studies that have specifically attempted to define and study members of the underclass. Each of these studies has stated that the underclass is multiracial. It is curious, therefore, that most of the studies that focus on the black underclass rarely address whether its members are inherently different from other members of the underclass, or if the characteristics of all individuals in the underclass are essentially the same but the factors that lead to their membership in the group are different.

In general, recent writings on black members of the underclass have used aggregate data based on the black population to document the social problems that are hypothesized to be caused by members of the group. Specifically, statistics that illuminate both the alarming number of black teenage pregnancies as well as the growing number of crimes committed by black men are often cited as proof of the social problems that are believed to be disproportionately caused by black members of the underclass. Few persons would deny that these social problems are among the most important issues facing the black community. The question to be addressed here is, do aggregate statistics provide sufficient evidence to document either the existence of, or problems caused by, members of the underclass?

If members of the underclass are defined as those individuals who have attitudinal and behavioral deficiencies, then the answer is no. Data are needed that link individuals who are defined as being members of the underclass (i.e., who are poor and have attitudinal and behavioral deficiencies) as disproportionately contributing to the specific social ills in question. As Richard Nathan and Kenneth Clark observe:

> The tendency is to conduct studies that analyze certain characteristics associated with underclass status, rather than try to group these characteristics for definitional purposes and then study the individuals in a particular [geographical] area who fit the definition. This is an important point. If the underclass group is one that should be

[†]*Discussants*: Bernard Anderson, Princeton University; Lynn C. Burbridge, The Urban Institute.

*Analyst, Education and Public Welfare Division, Congressional Research Service, Library of Congress, Washington, D.C. 20540. This paper represents my views and not those of the Congressional Research Service.

> defined in terms of multiple characteristics, we need data about the people who have such characteristics.
> [1982, p. 34]

Without causal data, it is impossible to directly link members of the underclass to the varying social problems for which they are hypothesized to be responsible.

Despite the lack of a uniform definition or substantial empirical evidence, policy discussions have already begun on how members of the black underclass might be aided. One strategy suggests that community based self-help programs may be the only effective way of changing the values and attitudes of the black underclass so that its members may take advantage of existing opportunities (see Glenn Loury, 1984). There is no question that community-based self-help programs aimed at improving the attitudes of those in the underclass are desirable and should be encouraged. At issue here is whether self-help programs should be viewed as augmenting or replacing government efforts. The view to be taken here is: 1) there exists a continuing need for antidiscriminatory government efforts; 2) without such efforts it is likely that more blacks will enter the underclass; and 3) the long-run success of self-help strategies may depend, in large part, on an environment that insures equal opportunity and access.

II. A Three-Dimensional Definition[1]

A criminal underclass subgroup was identified in the National Supported Work Demonstration set by using the 27-month ex-offender cohort that satisfied three criteria or dimensions. These dimensions are: 1) economic, have low income; 2) behavioral, have engaged in some deviant activity within the last year; and 3) attitudinal, exhibit some type of asocial attitude toward work. The objective of the demonstration, which used random assignment to create control and experimental groups, was to prepare individuals with severe labor market difficulties to find unsubsidized employment. (See Manpower Demonstration Research Corporation, 1980.)

The theoretical argument to be advanced here posits that the underclass is not a homogeneous group. On the contrary, the underclass is composed of subgroups that have different behavioral and attitudinal deficiencies. An underclass subgroup is defined as any group of economically disadvantaged individuals who display a common deviant behavior, and who also possess specific deviant attitudes with respect to the behavior they display. The difference between past definitions of the underclass and the one proposed here is that the three-dimensional definition asserts that the specific attitudes and behavior that characterize a particular underclass subgroup are different.

Ex-offenders in the supported work sample met both the economic and behavioral criteria due to the program's eligibility requirements of limited income and recent incarceration. The attitudinal scale used to distinguish underclass from nonunderclass ex-offenders is derived from the respondents' responses to 20 baseline questions which asked them to rate 10 legal and 10 illegal occupations from 0 to 100 in terms of the amount of respect they had for each. The 10 legal occupations are: house painter, postal worker or mail carrier, factory worker, prison guard, teacher, construction worker, policeman or policewoman, car washer, doctor, and cleaning person. The ten illegal occupations are: numbers runner, loan shark, hustler, cocaine dealer, purse snatcher, gambler, prostitute, counterfeiter, numbers banker, and numbers player or pimp. The attitudinal scale divides the average of the legal scores chosen by each individual by the average of the illegal scores. The scale has the desirable property that as a ratio, it measures the relative difference that an individual perceives between legal and illegal activities. It can be interpreted as indicating how much more a person values legal occupations over illegal ones.[2]

[1] For a more detailed explanation, see my dissertation (1985).

[2] If the mean illegal score of any respondent is less than one, his mean illegal score is set at one. This avoids the problem of dividing the mean legal score by zero which is an undefined solution.

Ex-offenders were classified as members of the underclass if they had attitudinal scores that were between 0 and 2 (lowest two quintiles); the borderline group if their score was greater than 2 and below 3.47 (middle quintile); and the nonunderclass if they were above 3.47 (highest two quintiles). It was believed that if differences in labor market behavior and experiences did exist based on attitudes toward legality, they could be discovered by comparing the upper two-fifths of the sample to the lowest two-fifths.

In addition to several factors which may have introduced some bias into the responses, there are also at least two conceptual problems with this attitudinal scale. First, the scale does not directly take into account the availability of earning income through work as opposed to criminal activity. The underlying assumption in using the scale is that individuals who give more respect to illegal over legal occupations are most likely people who would continue to engage in criminal activity regardless of the job opportunities that might be available. The second conceptual problem stems from the inability of this research to determine the causal direction between attitudes and behavior. Even if the assumptions underlying the attitudinal scale are true, and there is no evidence that they are, there is no way to determine whether the attitudes in question are the result or the cause of illegal activity.

TABLE 1—ESTIMATED COEFFICIENTS USED TO PREDICT SEMIMONTHLY EARNINGS OF EX-OFFENDERS FOR EACH 9-MONTH INTERVIEW PERIOD

	9-Month	18-Month	27-Month
Age (Over 30 years)			
21 Years or Less	4.41	23.74	8.47
	(27.63)	(51.71)	(56.92)
21–30 Years	20.33	23.89	−50.00
	(23.98)	(44.11)	(49.40)
Race (White)			
Nonwhite	−47.98[c]	−144.04[a]	−107.63[c]
	(32.50)	(59.72)	(66.95)
Sex (Male)			
Female	−48.36	−79.39	−98.43
	(38.15)	(70.12)	(78.59)
Education (Less than H.S. Diploma)			
H.S. Diploma or higher	68.21[a]	127.00[a]	74.74[b]
	(18.05)	(33.20)	(37.18)
Attitudinal Classification (Underclass)			
Nonunderclass	85.44[a]	109.76[b]	102.06[b]
	(27.87)	(51.45)	(57.41)
Borderline	54.23[b]	75.88	−2.57
	(32.03)	(58.79)	(65.97)
Where More $ Can be Made (Same on both)			
Street	−8.03	−36.63	−25.09
	(26.98)	(49.57)	(55.58)
Job	−31.06	−47.28	32.16
	(28.80)	(52.99)	(59.33)
Program Status (Control)			
Experimental	−75.64[a]	−33.52	−11.20
	(24.18)	(44.34)	(49.80)
Program Effect on Subgroups (Underclass)			
Nonunderclass	−59.80[c]	−43.91	−59.12
	(38.97)	(71.72)	(80.29)
Borderline	−27.16	−46.12	38.62
	(44.33)	(81.71)	(91.32)
Constant	124.30[a]	262.92[a]	336.20[a]
R^2	.141	.077	.048
Number of Cases	610	610	610

Note: Excluded categories and standard errors are shown in parentheses.

[a]Significant at $p = .01$ using a one-tailed test.
[b]Significant at $p = .05$ using a one-tailed test.
[c]Significant at $p = .10$ using a one-tailed test.

III. Empirical Findings

Table 1 shows the estimated regression coefficients for predicting the average semimonthly earnings of ex-offenders during the 9, 18, and 27-month interview periods. The independent variables are age, race, sex, education, attitudinal classification, "where more money can be made," supported work program status, and effect of supported work on program status. These variables are the baseline responses of the respondents prior to entering the demonstration. The variables education, race, and nonunderclass status are statistically significant across the three interview periods. Of particular interest is that having a nonunderclass attitude is expected to increase the individual's semimonthly earnings by about $100 over that of those who have underclass attitudes. This indicates that an individual's attitude has a significant influence on his or her earnings in the labor market. As noted above, the direction of causality between attitudes and income cannot be determined.

There are several interpretations that might be given to explain the consistently negative relationship that is shown between earnings and being nonwhite. The most accepted explanation is racial discrimination (see David Swinton and Lynn Burbridge, 1981). It is interesting to note that the decrease in earnings due to being nonwhite is almost exactly offset by the increase in earnings due to being nonunderclass. Stated differently, hold-

ing all other variables constant, nonunderclass minorities are expected to earn as much as underclass whites. These results indicate that minority members of the underclass have a much more difficult time in the labor market than white members of the underclass, and suggest that minority members of the *nonunderclass* are treated the same as white members of the *underclass* in terms of the earnings they receive in the labor market, all other characteristics being equal.

The findings above suggest that discrimination is equally as important as individual attitudes in limiting the economic success of nonwhites, at least with regard to ex-offenders. In short, there appears to be a continuing need for antidiscriminatory government policies. This leads to another consideration. It has been suggested that discrimination is a key factor in the development of underclass attitudes among blacks (see Douglas Glasgow, 1980). If this is true, self-help strategies that ignore this causal relationship could be expected to meet with little success without continued government efforts to eliminate discrimination.

Several variables that are not statistically significant are also revealing. The supported work program had no effect in increasing the earnings of either the ex-offender cohort in general (experimental/control group analysis), or its underclass and nonunderclass subgroups. These findings suggest that preparing members of the criminal underclass for jobs in the regular labor market using traditional employment and training programs would be a difficult and expensive task. However, before any government programs aimed at directly aiding members of the underclass can be considered, our basic understanding of the dynamics of the group should be substantially increased.

IV. Conclusion

There is very little that is known about the American underclass. There is even less that is known about different underclass subgroups or the experiences of different racial and ethnic groups within those subgroups. It is clear that aggregate data are inappropriate for making generalizations about members of the underclass when they are defined using attitudinal and behavioral criteria. This analysis has provided some limited evidence that individual attitudes do play a role in influencing income over time. Similarly, there is some evidence that discrimination continues to retard the economic advancement of racial minorities. Taken together, these preliminary findings suggest that self-help strategies aimed at aiding black members of the underclass that are presented in such a way as to limit government's role in reducing discrimination or promoting full employment may be premature.

REFERENCES

Auletta, Ken, *The Underclass*, New York: Random House, 1981.

Carson, Emmett D., "A Quantitative Analysis of The Underclass Concept," unpublished doctoral dissertation, Princeton University, October 1985.

Glasgow, Douglas, *The Black Underclass*, New York: Vintage Books, 1980.

Loury, Glenn C., "Internally Directed Action for Black Community Development: The Next Frontier for "The Movement,"" *The Review of Black Political Economy* Summer/Fall 1984, 31–46.

Nathan, Richard P. and Clark, Kenneth, "The Urban Underclass," in *Critical Issues for National Urban Policy: A Reconnaissance and Agenda for Further Study*, National Research Council, Washington: National Academy Press, 1980, 33–46.

Swinton, David H. and Burbridge, Lynn C., "Civil Rights and The Underclass," Washington: Urban Institute, October 1981.

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration*, Cambridge: Ballinger, 1980.

Transfer Payments, Sample Selection, and Male Black-White Earnings Differences

By WAYNE VROMAN*

In the period since the passage of the 1964 Civil Rights Act, the relative position of blacks in U.S. society has changed in several ways. Indicators of mortality, educational attainment, occupational status, and earnings all show large gains for blacks relative to whites. At the same time, however, blacks continue to experience severe disadvantages vis-à-vis whites in such areas as family stability, unemployment rates, average income, poverty rates (particularly among children), and dependence on government transfer payment programs. From the diverse statistical indicators of relative status, two sharply conflicting interpretations of black experiences since 1964 have emerged. One stresses relative improvements and convergence towards a position of parity with whites. The second stresses the lack of overall progress. A variant on the second point of view emphasizes a growing polarization within the black community; some succeeding in an increasingly demanding technological society while others who fail to succeed remain mired in dead-end jobs, the underground economy, and/or a position of dependence on government transfer payment programs.

The present paper examines the relative earnings of black men and closely related issues of labor force participation and the receipt of transfer payments. Although people with different perspectives can disagree on the extent of the gains realized by blacks since 1964, everyone would agree that a key requirement for sustained long-term improvement is the movement of black male earnings towards parity with white workers.

Among the possible indicators of relative earnings one frequently used measure is the black-white ratio of median annual earnings. An examination of such ratios using time-series data available from the U.S. Census Bureau and from the Social Security Administration shows three main "facts" about black men's relative earnings. 1) Relative earnings showed no important upward trend in the 20 years before 1965. 2) Between the mid-1960's and the mid-1970's, all series showed the black-white ratio to increase by about 10 percentage points from roughly .55 to about .65. 3) Since the mid-1970's, black men's relative earnings have not shown an important continuing trend towards earnings parity with whites. (See Part I of my 1985 paper.)

Theories to explain the racial earnings gap typically stress either supply or demand side factors as being of primary importance (see Ray Marshall, 1974). A particularly interesting explanation for the post-1964 gains has been offered by Richard Butler and James Heckman (1977). They argue that growth in government transfer payment programs since 1964 has been responsible for an apparent gain in black relative earnings. Their argument emphasizes the high replacement rates that transfers represent for low-wage workers. Since blacks usually earn much less than whites, disproportionate numbers of black workers would be induced to stop working by the availability of transfers. They argue that transfer-induced labor supply reductions have caused the published earnings medians (based on workers with reported earnings) to be increasingly affected by a problem of sample selection. Transfers induce a labor supply response, remove low-wage workers from the earnings distribution, artificially inflate the published earnings medians, and have a larger effect on the medians for black men.

Several testable hypotheses are implied by the preceding reasoning. 1) Black male labor supply reductions would be larger than the white male reductions. 2) Labor supply

*The Urban Institute, 2100 M Street, NW, Washington, D.C. 20037. Financial support for my research was provided under NSF grant no. SEC-830-9698.

reductions would be associated with an increased receipt of transfer payments. 3) Those who stop working in order to receive transfers would be drawn disproportionately from the lower tail of the black male earnings distribution. Each of these topics has been examined in my 1985 paper.

I. Labor Supply Reductions

A key element in the Butler-Heckman hypothesis is that labor supply reductions occur in the form of complete labor force withdrawal. Reductions in hours of work would cause earnings medians to decline and give accurate information on relative earnings. For movement in the median to give a misleading signal, it is necessary for low-wage workers to withdraw completely from the labor force.

Information on persons who did not work at all during the year is available from the annual work experience surveys conducted by the Census Bureau for the U.S. Labor Department. Proportions of the population who did not work were traced from 1958 to 1984 for white and nonwhite men of various ages. (Data for blacks are available only since 1975.) In nearly all age-race groups, the proportion of men with zero earnings has grown. The largest changes occurred among older men of both races (ages 45–54, 55–59, 60–64, and 65 and older) and among young nonwhite men aged 16–19 and 20–24. Important trends in male labor supply reductions were apparent. In each age group, the labor supply reductions were larger for nonwhites than for whites, but the biggest contrasts were found among those aged 16–19 and 20–24. For example, between 1958 and 1984, the zero earnings proportions among 20–24-year-olds increased from .112 to .282 for nonwhite men while it actually decreased modestly (from .107 to .089) for white men.

To investigate more systematically the racial trends in nonwork status, multiple regressions were fitted for the 1958–83 period. The dependent variable was a fixed-weight index of age-specific zero earner proportions for men of each race. Explanatory variables were a cyclical control (the unemployment rate for men 35–54), a long-term trend and a trend that operated just from 1964 to 1973 (the years when the black-white median earnings ratio showed a clear upward trend). Three findings were most important. First, the nonwhite zero earner proportion displayed greater cyclical sensitivity than did the white proportion. Second, the long-run upward trend in zero earnings status was much larger for nonwhites than for whites. Point estimates of the annual increases were .006 and .003 for nonwhites and whites, respectively. Third, no evidence of an accelerated trend toward zero earner status during the 1964–73 period was found for men of either race. The regressions using work experience data provided no support for the hypothesis that labor force withdrawal was unusually rapid between 1964 and 1973.

II. Increased Receipt of Transfer Payments

Labor supply reductions and withdrawal from the labor force are frequently associated with an increased utilization of transfer payments. The receipt of transfers by men of various ages was examined in two types of data; program data from the individual transfer programs and data from the *Current Population Survey* (*CPS*). Since 1968, the *CPS* has information on seven major groups of transfer programs which allow comparisons of recipiency rates by race and age. The seven are public assistance, unemployment insurance, workers' compensation, veterans' benefits, OASDHI (or Social Security), government pensions, and private pensions.

Overall, a measurable proportion of men receive transfer payments. In 1980, for example, about 30 percent of both nonwhite and white men age 16 and older received transfers sometime during the year. The highest beneficiary proportions for men of both races were found among men 65 and older. Among persons younger than age 65, the highest proportions were found among those age 45–54, 55–59, and, especially, 60–64. Nonwhite men 16–19 and 20–24 did not receive transfers in noticeably large and/or growing proportions between 1969 and 1983.

Summary comments about the individual transfer programs are as follows:

Public Assistance, Workers' Compensation, and Veterans' Benefits. Because these three programs have not shown an important

tendency to increase in size, that is, the number of beneficiaries relative to the total male population, they could not have caused a disproportionate number of labor force dropouts among nonwhite men during the 1964–73 period.

Unemployment Insurance (*UI*). Although nonwhite men are more likely to receive these benefits than are white men, this is caused by their higher unemployment rates and not because unemployed nonwhites have higher benefit recipiency rates than unemployed whites. Between 1964 and 1973, the nonwhite beneficiary proportions could not have increased much more rapidly than the white proportions. Given the impossibility of long-term benefit receipt by *UI* beneficiaries, this program is not a good candidate on a priori basis for causing disproportionate sample selection problems in the nonwhite earnings distribution.

OASDHI (Social Security). This program grew rapidly between 1964 and 1973 and the proportion of nonwhite men who were disability insurance (*DI*) recipients grew more rapidly than did the proportion for white men. However, there was a faster rate of growth in the nonwhite proportion between 1957 (when *DI* was instituted) and 1964, a period when relative earnings did not change. It is apparent in program data that the racial differences in the growth rates of the beneficiary proportions were of about the same size in the 1957–64 and 1964–73 periods.

Government Pensions and Private Pensions. These programs have grown more rapidly since 1972 than in early years, but when the two are considered together the net difference by race was not too great. Private pensions are received by proportionately more older white men while proportionately more older nonwhite men receive government pensions. The growth of these programs during 1964–73 was not of a scale sufficient to cause a much larger sample selection problem in the nonwhite male earnings distribution.

Thus, although adult men became increasingly likely to receive transfer payments between 1950 and the mid-1980's, and the nonwhite beneficiary proportion has increased more rapidly than the white proportion, contrasts by race during the 1964–73 period were not sufficiently large to cause a serious sample selection problem as hypothesized by Butler and Heckman. After examining transfer data by program, the one that stands as the best candidate for causing sample selection problems in that particular time period is neither public assistance nor unemployment insurance (as they suggested) but rather Social Security *DI*. Even with the *DI* program, however, the fact of its more rapid growth among nonwhites in the 1957–64 period (when the median earnings ratio was stable) stands inconveniently at odds with their hypothesis.

III. The Prior Earnings of Transfer Recipients

To investigate the prior earnings of current transfer recipients one needs longitudinal data. A data file present at the Urban Institute combines micro data from the March 1978 *CPS* with earnings histories from the Social Security Administration's Summary Earnings Record (*SER*). The *CPS* provides information on work status and receipt of transfers (by program) during 1977 while the *SER* has annual data on Social Security covered earnings for each year between 1950 and 1980. With this data base, one can examine the prior earnings of persons who were both transfer recipients and nonworkers in 1977. Comparison data were taken from an earlier study that also investigated racial earnings differences using Social Security earnings data (see my 1974 paper).

The analysis can be summarized with a few statements. The 1977 transfer recipients who stopped working were mainly older men. This agrees with earlier findings that the receipt of transfers increases with age. For men 25–44 and 45–64, the median earnings of subsequent transfer recipients were somewhat (5 to 10 percent) lower than the medians for other men of the same age. Because age-earnings profiles slope sharply upward at younger ages, however, these medians were at least 10 percent higher than the overall medians for all workers 16–64. The higher-than-average medians were found for both black and white male transfer recipients. Thus, the prior earnings of subsequent trans-

fer recipients who had stopped working were not unusually low.

Within the group of 1977 transfer recipients, it was particularly instructive to examine the prior earnings of men 25–44 in 1957. One could argue that without the availability of transfers these men who were younger than 65 in 1977 would have still been working in 1977. For black and white men in this age group, the 1957 earnings medians were, respectively, $2,510 and above $4,200. These medians were more than 10 percent higher than median for all men 16–64. Their dropout behavior tended to lower the overall median just as did the dropout behavior of other subsequent transfer recipients.

IV. Conclusions

To summarize, three statements can be offered. 1) Although there has been a bigger trend in labor supply reductions among black men, no discernable acceleration in the trend occurred during 1964–73, the period of improvement in the black-white median earnings ratio. 2) There has been a trend towards increased receipt of transfers among men of both races, but no unusually large trend was observed for black men between 1964 and 1973. 3) The previous earnings median for 1977 transfer recipients (of both races) who had stopped working was higher (not lower) than the median for all workers. Thus the transfer payments-sample selection explanation for increases in the black-white earnings ratio, though an interesting hypothesis, is not supported when its key elements are subjected to empirical testing.

REFERENCES

Butler, Richard and Heckman, James, "The Government's Impact on the Labor Market Status of Black Americans: A Critical Review," in Leonard Hausman et al., eds., *Equal Rights and Industrial Relations*, Madison: Industrial Relations Research Association, 1977, 235–81.

Marshall, Ray, "The Economics of Racial Discrimination: A Survey," *Journal of Economic Literature*, September 1974, *12*, 849–71.

Vroman, Wayne, "Changes in Black Workers' Relative Earnings: Evidence from the 1960's," in George von Furstenberg et al., eds., *Patterns of Racial Discrimination*, Vol. 2, Lexington: Lexington Books, 1974, 147–96.

______, "The Relative Earnings of Black Men: Theories and Evidence," Urban Institute, December 1985.

Federal Courts and the Enforcement of Title VII

By JEROME MCCRISTAL CULP, JR.*

Twenty years of Title VII enforcement has sparked a spirited debate concerning the effectiveness of the 1964 Civil Rights Act in improving the labor market experience of black Americans. Some have suggested that government efforts to eliminate discrimination had little to do with the post-1964 rise in the relative incomes of black Americans. Most investigators have concluded that since the passage of Title VII, the income of blacks relative to whites has increased, that this increase was greatest for young blacks and for black women relative to white women, and that government action contributed to the improvement in the economic position of black Americans (see Charles Brown, 1984).

Few investigators have found that the improvement in the economic position of black Americans can be directly traced to the enforcement of Title VII. These investigations have in general ignored the role that Title VII plays in the marketplace and the role the courts have played in ending discrimination. The few attempts to directly estimate the impact of courts have found that favorable court decisions are associated with greater black improvement (see Paul Burstein, 1979). Even these limited efforts are likely to be an underestimate of the impact of the courts on black improvement because the threat of litigation will also alter behavior.

Congress decided when it passed Title VII of the 1964 Civil Rights Act to limit the enforcement powers of the Equal Employment Opportunity Commission, and to emphasize conciliation and settlement of disputes between employees and their employers. These twin choices have made government efforts to reduce racial discrimination a product of the actions of the federal courts and individual plantiffs. Court decisions which find employers guilty of racial discrimination are likely to increase the apparent costs to employers of engaging in discriminatory behavior. The way in which courts function explains some of the surprising changes in the relative position of black Americans during the last twenty years.

This paper has two purposes. First, I will show that a simple model of the regulation of discrimination by the courts will provide a better view of some puzzling aspects of recent black employment history. Second, I will show that the courts have helped black Americans secure greater economic equality.

I. The Courts and Antidiscrimination Enforcement

The relative income of black males and females has risen significantly since 1964. This rise in income levels and occupational position was faster than the pre-1964 period for black men. The gains declined for black men after 1978 and the upward trend for black women stopped. Younger cohorts of black youth have done better than older cohorts and the college educated have done better than high school graduates. The difference in income between younger black men and younger white men is smaller than the difference in income between older black men and older white men.

Traditional explanations have attributed these different rates of progress of blacks to differences in cohort quality (see James Smith and Finis Welch, 1977) or the human capital cost of taking advantage of these improvements (see Richard Freeman, 1981). Neither explanation is entirely satisfactory. It is difficult to argue that increases in black incomes can be attributed to increasing education levels of black youths when most have been trained in poor urban schools. It is hard to see why new entrants ought to be at an advantage against more experienced black cohorts who have been trapped in poor jobs. Since these cohorts have little to lose in

*Associate Professor of Law, Duke University Law School, Durham, NC 27706. I thank Bernard Anderson for helpful comments.

terms of wages or human capital investments, one might expect them to be the greatest beneficiaries of affirmative action. This is especially true since the Supreme Court requires blacks to individually prove that they suffered from past discrimination in order to successfully bring a Title VII action. In addition, this does not explain why the college educated do better than high school graduates. According to the human capital theory, all young workers ought to have similar levels of investment in job-specific human capital. The way in which courts enforce Title VII helps to explain these facts.

One reason why the activities of the courts have been ignored is the lack of a consensus on how the courts can act as economic regulators of activity. Richard Posner (1972) has offered one way of viewing the activity of the courts. Posner argues that judges and juries have a tendency to choose the efficient (or, as he later modified the terminology, the wealth maximizing) rule. Posner argues that the common law (judge-made law) system in the long run will choose the rule that leads to the greatest wealth.

The courts more than any other institution of government regulation retain aspects of the market. Unlike an administrative agency or an executive department, the court can enforce only those cases and actions that are brought to it by disgruntled litigants. The courts of course are used by executive agencies (particularly the EEOC and before that the Department of Justice) to eliminate patterns of discrimination. However, most cases are not brought by the government, but by disgruntled individuals seeking redress.

The supply of Title VII cases will be a function of the level of discrimination, the cost of bringing the case to court in terms of attorney fees and other litigation costs, and the nonlitigation costs to the individual of bringing the case to court. The latter costs include the lost goodwill of the employer which may show up in a higher probability of discharge, layoff, and reduction in hours, and a lower probability of promotion, and merit and other kinds of wage and nonwage increases, and the ability to change employers. The nonlitigation costs will be lower the less an employee has invested in a particular job and employer. This explains why Title VII litigants tend to be of two types. The first are disgruntled present employees who have decided to leave and often have secured another job or returned to school. These litigants are usually so disappointed that they will quit whether they win or lose their cases. The second group of litigants is disappointed job seekers. These individuals often have little to lose because it is generally impossible for a nonemployer to influence future job prospects through a bad reference. For current employees, references are important.

These nonlitigation costs explain why young people and black women have made more progress than older black men. Young people are more likely to fall into the second group of complaintants, that is, disgruntled applicants. These young people also have less to lose from bringing suits against an employer than would older employees because it is easier for them to explain why they left a job and returned to school or childrearing. These nonlitigation costs also explain why the well educated have done better than those with less education despite the fact that courts have been reluctant to intervene in the employment processes of jobs that require higher levels of education. The well educated are more likely to make the investment in documentation required to succeed in a Title VII case.

Thus, the federal courts at the beginning of the enforcement of the antidiscrimination provisions of Title VII adopted a set of procedural rules which were primarily group centered and substantive. In *Griggs v. Duke Power Co.* (1971), the Supreme Court found that an employer's practice that was otherwise neutral violated Title VII because neutral practices exclude most blacks. *Griggs* made use of employment procedures that had a disproportionate impact on black workers or applicants, and was proven unlawful. Courts, however, have over time reduced the applicability of these group practices and limited Title VII protection to procedural rights of individuals. In recent

years this has resulted in courts being less effective in altering the behavior of employers (see my 1984 paper).

There are several reasons federal courts are likely to set the level of discrimination below its current level, but above zero discrimination. First, judges are not trained in economic theory, econometrics, or labor market analysis. It is therefore difficult for judges to ascertain improvement in the labor market experience of blacks relative to whites. If judges are able to discern the improvement in the condition of blacks, the judge may not be able to determine if the improvement was permanent or which factors lead to black improvement. If judges are wrong about the labor market experience of blacks, then their conclusions about the direction of equality need not be correct. Second, the court uses procedural rules to effectuate its policies. If a judge can tell whether there has been black improvement, it will still be difficult for judges to know how these procedural rules translate directly into black economic improvement. However, the courts have helped to end racial bias.

II. Federal Courts and Police and Fire Departments

Economists generally have difficulty finding theoretical basis for market discrimination. In a competitive market, there are pressures for nondiscriminators to drive out discriminators. Despite these large pressures against market discrimination, there is still substantial evidence that racial discrimination exists in the American labor market. One explanation for the continued vitality of racial discrimination is that not all labor markets are competitive. Economists have known for a long time that in markets where employers have monopsony power, discrimination can persist. In commenting on the role of courts in ending racial discrimination, economists have not been careful in explaining that, even if market pressures will eliminate racial discrimination in competitive situations, many of the circumstances where racial discrimination is charged are situations where employers have monopsony power.

TABLE 1—PERCENTAGE CHANGES IN POLICE AND FIRE DEPARTMENTS IN TWENTY AMERICAN CITIES, 1970–80[a]

	Fire	Police
Black Males	+134.	+22.
White Males	−15.	−23.
Total Males	−2.	−19.
Total Females	+32.	+79.
Total	−1.	−16.

[a] The cities are New York, Chicago, Los Angeles, Philadelphia, Houston, Detroit, Dallas, San Diego, Phoenix, Baltimore, San Antonio, Indianapolis, San Francisco, Memphis, Washington, D.C., San Jose, Milwaukee, Cleveland, Columbus, Boston (by 1980 Census).

The two most famous Title VII cases decided by the Supreme Court fall clearly into this category. Both the *Griggs* and the *United Steelworkers v. Weber* (1979) cases involve labor markets where the employers were large in relatively isolated communities. Both employers have some monopoly power in their product markets. Market pressures are unlikely to alter such employer behavior without the intervention of some outside force. A substantial number of the labor markets covered by Title VII are markets in which employers have some monopsony power. One market covered by Title VII where monopsony power exists is the labor market for police and fire departments. This market is interesting because the Supreme Court has decided or will decide several cases involving discrimination in that market. Both cases raise the question of what is the proper level of discrimination and who should bear the burden of past discrimination and present economic fluctuations.

Table 1 describes employment changes in police and fire departments in the twenty largest cities in the country during 1970–80 for several demographic groups. During the 1970's, the total number of police and fire departments remained approximately the same. However, the number of black and female police and fire officers increased dramatically. This change is primarily attributable to the existence or threat of court action. Much of this action was initiated by the Justice Department, but individual com-

plaintants clearly contributed to the changing complexion of these departments.

Those cities with consent decrees show a marked improvement in the number of black and female officers. It is interesting to note two other consequences in the one city (Philadelphia) that did not have a consent decree for race but did have one for sex—the sexual makeup of the department changed but not the racial makeup. After 1974 when Title VII was made applicable to local government police and fire departments, these departments showed some upward movement even if consent decrees were not entered. Fire and police departments have monopsony power and in the absence of court or other pressures job prospects in this very important area are unlikely to change. In addition, if courts alter their view of the effectiveness of Title VII as they seem to be, little progress for blacks can be expected in these monopsony markets in the near future.

REFERENCES

Anderson, Bernard E. and Wallace, Phyllis A., "Public Policy and Black Economic Progress: A Review of the Evidence," *American Economic Review Proceedings*, May 1975, *65*, 47–52.

Brown, Charles, "Black/White Earning Ratios Since the Civil Rights Act of 1964: The Importance of Labor Markets Drop Outs," in R. Ehrenburg, ed., *Research in Labor Economics*, New York: JAI Press, 1984.

Burstein, Paul, "Equal Employment Opportunity Legislation and the Income of Women and Non-Whites," *American Sociological Review*, June 1979, *44*, 367–91.

Culp, Jerome, "A New Employment Policy for the 1980's: Learning from the Victories and Defeats of 20 Years of Title VII," mimeo., 1984.

Freeman, Richard B., "Black Economic Progress after 1964: Who has Gained and Why?," in Sherwin Rowen, ed., *Studies in Labor Markets*, Universities-National Bureau Conference Series, No. 31, Chicago: University of Chicago Press, 1981.

Posner, Richard, *The Economic Analysis of the Law*, Boston: Little Brown, 1972.

Smith, James P. and Welch, Finis, "Black-White Wage Ratios: 1960–70," *American Economic Review*, June 1977, *67*, 323–38.

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

United Steelworkers v. Weber, 444 U.S. 889 (1979).

What was Affirmative Action?

By JONATHAN S. LEONARD*

In recent years there have been two major criticisms of affirmative action. The first is that affirmative action does not work; therefore, we should dispose of it. The second is that affirmative action does work; therefore, we should dispose of it. My concern here will be with the first of these criticisms. If affirmative action has not changed the employment patterns of nonwhites and females, then much of the discussion since 1965 of its philosophical merits amounts to shadowboxing. Has affirmative action been successful in increasing employment opportunities for females and minorities?

The literature on affirmative action can be divided into studies of the regulatory process that find it mortally flawed and studies of impact that find it successful. In light of the unanimity of the process studies in finding the affirmative action regulatory mechanism seriously deficient, it is surprising that the few econometric studies of the impact of affirmative action in its first years (see Charles Brown, 1984, for a review; Orley Ashenfelter and James Heckman, 1976, for an example), all based on a comparison of EEO–1 forms by contractor status, have generally found significant evidence that it has been effective for black males. These few studies of the initial years of affirmative action (1966–73) are not directly comparable because of different specifications, samples, and periods. They do find, nevertheless, that despite weak enforcement in its early years, and despite the ineffectiveness of compliance reviews, affirmative action has been effective in increasing black male employment share in the contractor sector, but generally ineffective for other protected groups.

*Assistant Professor of Industrial Relations, School of Business, University of California, Berkeley, CA 94720, and NBER-Olin Fellow.

I. The Impact of Affirmative Action on Employment

Affirmative action under the executive order applies only to federal contractors. One method of judging its effect then is to compare the growth of minority and female employment at federal contractor establishments with their employment growth at similar noncontractor establishments. With the cooperation of the U.S. Department of Labor, I performed such a comparison using employment data reported in 1974 and 1980 by 68,690 establishments with more than 16 million employees.

The results of this study indicate that affirmative action has been far more effective than its critics allow. Between 1974 and 1980, black male and female, and white female employment shares increased significantly faster in contractor establishments subject to affirmative action than in noncontractor establishments. The other side of this coin is that white males' employment share declined significantly more among contractors. The growth rate is 3.8 percent faster for black males, 7.9 percent for other minority males, 2.8 percent for white females, and 12.3 percent for black females. White male employment grew 1.2 percent slower in the contractor sector (see my 1984a article).

Compliance reviews, the major enforcement tool, have played a significant role over and above that of simply being a contractor, advancing black males by 7.9 percent, other minority males by 15.2 percent, and black females by 6.1 percent among reviewed establishments. Compliance reviews have retarded the employment growth of whites, and curiously may have reduced white female employment growth in the reviewed establishments. This anomalous result for white females is difficult to reconcile with the countervailing and dominant positive impact of contractor status on white

females, but may be influenced by a review process that asks for more than last year, rather than more than average, in a time of sharply increasing female labor supply. With the exception of white females, compliance reviews have an additional positive impact on protected group employment beyond the contractor effect. Direct pressure does make a difference.

Employment opportunities, and the efficacy of affirmative action, also depend critically on growth. Minorities and females experienced significantly greater increases in representation in establishments that were growing and so had many job openings. Other establishment characteristics also make a difference. Establishments that are not part of multiplant corporations have significantly lower growth rates of employment of members of protected groups. Corporate size is probably of greater consequence than establishment size, with larger corporations showing greater increases in minority and female employment. Establishments that are nonclerical white-collar intensive exhibit faster employment growth for both male and female blacks, and significantly slower growth for white males. Unionized establishments generally have comparable or better records than those without unions (see my 1985c article). Also, while the central focus of this analysis is on affirmative action under the executive order, it should be understood that the executive order has functioned within the backdrop of Title VII's congressional mandate and substantial legal sanctions.

II. Occupational Advance

One of the major affirmative battlefields lies in the white-collar and craft occupations. It is in these skilled positions that employers are most sensitive to productivity differences and have complained the most about the burden of goals for minority and female employment. It is also in these occupations that the potential wage gains to members of protected groups are the greatest.

The contract compliance program has raised the demand for black males more in the highly skilled white-collar and craft jobs than in the blue-collar operative, laborer, and service occupations. While this may help explain why highly skilled black males have been better off than their less-skilled brethren, it does not help explain why low-skilled black males should be having greater difficulty over the years in finding and holding jobs, because affirmative action has generally increased the demand for black males, even in unskilled jobs (see my 1984b article).

Affirmative action has also helped nonblack minority males. There is evidence of a twist in demand toward Hispanic, Asian, and American Indian males in white-collar occupations, particularly in sales and clerical positions, and away from this group in operative and laborer position. The evidence within occupations suggests that the contract compliance program has had a mixed and often negative impact on white females. In contrast to whites, black females in contractor establishments have increased their employment share in most occupations.

While part of this occupational upgrading may be overstated because of biased reporting, in particular the upward reclassification of minority or female intensive occupations (James Smith and Finis Welch, 1984), the finding of occupational advance is reinforced by evidence that affirmative action has narrowed the difference in earnings between the races by raising the occupational level of nonwhite males.

Similarly, affirmative action does not appear to have directly contributed to the economic bifurcation of the black community, but rather appears to increase the demand for lowly educated minority males as well as for the highly educated (see my 1984d paper).

If minorities and females do not share the skills and interests of white males, then perhaps the best one can expect from an affirmative action program is to increase their employment. But to the extent that minorities and females share the qualifications and interests of white males, an effective affirmative action program should also improve their chances of sharing the same occupations.

Just as no policy works in isolation, so none can be evaluated in isolation. The success of this program in skilled occupations is probably due in part to the increasing supply of skilled minorities in many fields,

as well as to the more aggressive use of sanctions after the early 1970's. The lesson drawn is that affirmative action programs work best when they are vigorously enforced, when they work with other policies that augment the skills of members of protected groups, and when they work with growing employers.

III. Goals or Quotas?

Have these employment advances been achieved through the use of rigid quotas? The goals and timetables for the employment of minorities and females drawn from federal contractors under affirmative action stand accused on two mutually inconsistent charges. The first is that "goal" is really just an expedient and polite word for inflexible quotas. The second is that these goals are worth less than the paper they are written on, because affirmative action has never been stringently enforced. Is negotiation over affirmative action goals an empty charade played with properly penciled forms, or does it in fact lead to more jobs for minorities and females in the contractor sector? If the latter is the case, are these goals so strictly adhered to as to constitute quotas?

Goals set in these costly negotiations do have a measurable and significant correlation with improvements in the employment of minorities and females at reviewed establishments. At the same time, these goals are not being fulfilled with the rigidity one would expect of quotas. While the projections of future employment of members of protected groups are inflated, (by roughly a factor of 10) they are not hollow: the establishments that promise to employ more do actually employ more (see my 1985b article).

We have a regulatory process that appears to be effective in its whole and ineffective in its parts. The paperwork requirements of the affirmative action plan, the notification and resolution of plan deficiencies, and even conciliation agreements and show-cause notices appear to have little significant impact on subsequent employment demographics when individually considered. Nevertheless, protected-group employment share does generally grow more rapidly at reviewed firms, and goals are strongly correlated with this growth. While much of the nit-picking over paperwork is ineffective, the system of affirmative action goals has played a significant role in improving employment opportunities for members of protected groups.

IV. The Targeting of Compliance Reviews

Affirmative action can be broadly conceived of as pursuing either antidiscrimination or job and earnings redistribution goals. It can either pursue equality or opportunity or equality of result, although given the historical record, progress toward one goal will often entail progress toward the other.

If one thought of the OFCCP's primary concern as fighting the most blatant forms of prima facie employment discrimination directly in the workplace, one might then expect reviews to be concentrated at establishments with a relative small proportion of females and black males, controlling for size, industry, and region. There is little consistent significant evidence of this in the past. In part, this may be explained by the requirement of pre-award compliance reviews. Establishments with the smallest proportion of minorities or females, are not consistently more likely to be reviewed. Reviews are significantly more likely to take place in nonclerical white-collar intensive establishments, and at both large and growing establishments where any costs to white males are likely to be more diffused (see my 1985a article).

In interviews, field officers of the OFCCP have stated that they do not generally look at an establishment's past employment record in targeting reviews. Reviewing large establishments with little regard for their past record of minority or female employment, as would seem to be required under the proposed new executive order, illustrates a lack of attention to attacking the grossest prima facie forms of current employment discrimination.

V. Antidiscrimination or Reverse Discrimination?

Despite poor targeting, affirmative action has helped promote the employment of

members of protected groups, and Title VII has likely played an even greater role. This raises the most important and the most controversial question: has this reduced discrimination, or has it gone beyond and induced reverse discrimination against white males? This is also the question on which our evidence is least conclusive. The finding of decreased employment growth for white males is not sufficient to answer the question since it is consistent with both possibilities.

The hypothesis inherent in some criticisms of affirmative action as reverse discrimination is that the relative marginal productivities of minorities and females have declined as their employment has increased, and have not moved toward equality with relative wages.

Using estimates of production functions relating output to inputs for the manufacturing sector, my 1984c article finds some weak evidence that relative minority and female productivity increased between 1966 and 1977, a period coinciding with government antidiscrimination policy to increase employment opportunities for members of these groups. No significant evidence is found here to support the contention that this increase in employment equity has had marked efficiency costs. Direct tests of the impact of governmental antidiscrimination and affirmative action regulation on productivity find no significant evidence of a productivity decline. These results suggest that antidiscrimination and affirmative action efforts have helped to reduce discrimination without yet inducing significant and substantial reverse discrimination. However, the available evidence is not yet strong enough to be compelling on either side of this issue. Since the productivity estimates are not measured with great precision, strong policy conclusions based on this particular result should be resisted.

VI. Conclusion

While numerical standards in the quest for equal opportunity may open the door to an emphasis on equal results, an affirmative action program without measurable results surely invites sham efforts. After all, executive orders barring discrimination by federal contractors have been with us since President Franklin D. Roosevelt, but these bore little if any fruit until policies relying on voluntary compliance were given teeth in the mid-1960's with a monitoring mechanism and a set of sanctions.

The evidence reviewed here is that a process that has been frequently criticized by various parties as either a system of draconian quotas or as an exercise in paper pushing has actually been of material importance in prompting companies to increase their employment of minorities and females. For a program lacking public consensus and vigorous consistent enforcement, this is a surprisingly strong showing. Hopefully, evidence of the effectiveness of past affirmative action programs will be of some use as we prepare to enter the second generation of policy by troubled euphemism: nonpreferential affirmative action.

REFERENCES

Ashenfelter, Orley and Heckman, James, "Measuring the Effect of an Antidiscrimination Program," in Orley Ashenfelter and James Blum, eds., *Evaluating the Labor Market Effects of Social Programs*, Princeton: Industrial Relations Section, 1976.

Brown, Charles, "The Federal Attack on Labor Market Discrimination: The Mouse that Roared?," in R. Ehrenburg, ed., *Research in Labor Economics*, New York: JAI Press, 1984.

Leonard, Jonathan S., (1984a) "The Impact of Affirmative Action on Employment," *Journal of Labor Economics*, October 1984, *2*, 439–63.

______, (1984b) "Employment and Occupational Advance under Affirmative Action," *Review of Economics and Statistics*, August 1984, *66*, 377–85.

______, (1984c) "Anti-Discrimination or Reverse Discrimination: The Impact of Changing Demographics, Title VII and Affirmative Action on Productivity," *Journal of Human Resources*, Spring 1984, *19*, 145–74.

______, (1984d) "Splitting Blacks? Affirma-

tive Action and Earnings Inequality Within and Between Races," NBER Working Paper No. 1327, April 1984.

______, (1985a) "Affirmative Action as Earnings Redistribution: The Targeting of Compliance Reviews," *Journal of Labor Economics*, July 1985, *3*, 363–84.

______, (1985b) "What Promises Are Worth: The Impact of Affirmative Action Goals," *Journal of Human Resources*, Winter 1985, *20*, 3–20.

______, (1985c) "The Effect of Unions on the Employment of Blacks, Hispanics, and Women," *Industrial and Labor Relations Review*, October 1985, *39*, 115–32.

Smith, James and Welch, Finis, "Affirmative Action and Labor Markets," *Journal of Labor Economics*, April 1984, *2*, 269–301.

Implementing Comparable Worth: A Survey of Recent Job Evaluation Studies

By ELAINE SORENSEN*

During the 1980's, equal pay for comparable worth has emerged as a major legislative issue, especially at the state and local level. In the wake of increasing implementation of comparable worth in the public sector, efforts have been made to estimate the probable effect of nationwide implementation of comparable worth-type legislation on the earnings gap between women and men (see, for example, George Johnson and Gary Solon, 1984). These analyses have generally concluded that a comprehensive comparable worth policy would have very little effect on the sex-based earnings gap. This paper offers alternative estimates of the effect of comparable worth on the male-female earnings gap, based upon examination of four state-level comparable worth studies (Iowa, Michigan, Minnesota and Washington).

I. Why Previous Estimates are Misleading

Johnson and Solon measure the effect of comparable worth policies on male and female earnings by estimating separate earnings equations for women and men with the proportion of women in an occupation, or F, as an independent variable, as well as other explanatory variables. Johnson and Solon claim that the goal of a comparable worth policy is to eliminate the estimated negative effect of the variable F on the earnings of women and men. They find that, if comparable worth were implemented in this manner, it would increase the earnings ratio between women and men by less than 10 percent. From this analysis, Johnson and Solon conclude that implementing comparable worth would have a very small effect on the male-female earnings gap.

The first problem with the Johnson-Solon analysis is that they do not restrict the benefits of comparable worth to workers in female-dominated occupations as do comparable worth studies. They claim that all workers would benefit from a comparable worth policy, except those employed in occupations that are exclusively male. Consequently, Johnson-Solon overestimate the effect of comparable worth on male earnings, and underestimate its effect on the earnings gap between women and men.

Second, Johnson-Solon and comparable worth studies use different dependent variables. The dependent variable in comparable worth studies is the occupational salary, defined as one of the salary levels that an individual could receive if employed in this occupation. On the other hand, Johnson-Solon use individual salaries of men and women as their dependent variable. But, individual female earnings and the occupational earnings of women are different.

Female earnings are affected by two inequalities, unequal pay within an occupation and unequal pay for comparable worth. Unequal pay within an occupation is negatively correlated with the proportion of women in an occupation. In other words, women in male-dominated occupations face greater earnings disparities within their occupation than women in female-dominated occupations. Because of this additional inequality and its negative correlation with the proportion of women in an occupation, female earnings are rather insensitive to the proportion of women in their occupation. Thus, the estimated negative coefficient on F in the

†*Discussants*: Sharon B. Megdal, University of Arizona; Marilyn Power, University of New Hampshire.

*Assistant Professor, Department of Economics, University of Massachusetts, Amherst, MA 01003.

female earnings equation is relatively small. However, its size is very important in Johnson and Solon's analysis, since they argue that a comparable worth policy would eliminate its effect on female earnings. Consequently, by using the personal earnings of women, Johnson and Solon allow the existence of unequal pay within an occupation to limit their measurement of unequal pay for comparable worth.

On the other hand, comparable worth studies do not allow their measurement of unequal pay for comparable worth to be affected by the existence of unequal pay within an occupation. Comparable worth studies examine occupational salaries, which, by definition, are not affected by unequal pay within an occupation. Consequently, the occupational earnings of women are more sensitive to the proportion of women in an occupation than female earnings. If Johnson and Solon had used women's occupational earnings, the estimated negative coefficient on F would have been larger. By using female earnings instead of women's occupational earnings, they underestimate the gains for women under a comparable worth policy, and underestimate its effect on the male-female earnings gap.

II. Estimating Comparable Worth's Effect from Job Evaluation Studies

According to advocates of comparable worth, the policy's purpose is to eliminate the effect of occupational segregation within a firm on the earnings disparity between women and men once legitimate factors, such as differences in job requirements, have been accounted for (Ronnie Steinberg, 1984). Most comparable worth studies measure the effect of occupational segregation on the male-female earnings gap by first estimating separate earnings equations for male- and female-dominated occupations.[1] These equations are

$$S_m = a_0 + a_1 J_m + u_m \tag{1}$$

$$S_f = b_0 + b_1 J_f + u_f \tag{2}$$

where m and f = male- and female-dominated occupations, respectively; S = the occupational salary for each occupation; J = the job evaluation score for each occupation; and u = the random error term.

Male- and female-dominated occupations are defined as any occupation in which 70 percent or more of the employees are male or female, respectively. This is referred to as the "70 percent rule" and has become the standard definition used in most comparable worth studies. The dependent variable is the occupational salary, defined as one of the salary levels that an individual could receive if employed in this occupation. Generally, the maximum salary level is used to represent the occupational salary, although entry and intermediate salary steps have also been used.

The job evaluation score is a number that measures the overall requirements of a job according to the criteria established by the job evaluation plan. All four states in this study selected an "a priori" factor-point plan to evaluate jobs, the most commonly used job evaluation plan in the United States. A set of factors and weights is selected before the commencement of the evaluation. The factors are expected to reflect the requirements of a job, and usually fall into four broad categories: skill, effort, responsibility, and working conditions. Weights are applied for each factor and indicate their relative importance. The Washington state study, for example, used the following categories in their analysis: "job knowledge," "mental demands," "accountability," and "working conditions." The weights were 47, 23, 27, and 3 percent, respectively. An evaluation committee evaluates jobs in terms of each factor and assigns a level of points commensurate with the amount of the factor required on the job. These factor scores are summed for each job to produce a total point score, or job evaluation score.

Comparable worth studies have focused upon occupational salaries and job require-

[1] Comparable worth studies have developed slightly different methods of measuring the extent of unequal pay for comparable worth. In order to compare across studies, I have adopted a set of procedures quite similar to those used by Ronald Ehrenberg and Robert Smith (1984).

ments, rather than individual salaries and human capital variables, because in the public sector, where these studies have taken place, occupational salaries are generally determined by a formal set of rules. These rules establish entry level salaries for each occupation and the incremental increases that individuals may achieve within that occupation. In addition, every occupation has a formal job description which specifies the basic requirements of the job, and it is presumed that any individual in the job meets these basic requirements.

A comparable worth policy can achieve its goal of eliminating the effect of occupational segregation on the sex-based earnings disparity by computing the earnings of female-dominated occupations from the estimated earnings equation of male-dominated occupations. This equation is

$$\hat{S}_f = \hat{a}_0 + \hat{a}_1 J_f \tag{3}$$

where f = female-dominated occupations; $\hat{a}_0$ and $\hat{a}_1$ = the estimated coefficients from the male-dominated occupational earnings equation; and J = the job evaluation score for each female-dominated job.

To determine the average percentage gain for workers in female-dominated occupations from a comparable worth policy, I calculated for each female-dominated occupation, the difference between the occupation's predicted earnings under a comparable worth policy and the occupation's current earnings, divided by the current occupational earnings (i.e., $(\hat{S}_f - S_f)/S_f$, where $\hat{S}_f$ is the predicted occupational salary under a comparable worth policy and S_f is the current occupational salary). This figure was calculated for each female-dominated occupation and weighted by the proportion of women in that occupation. These figures were then added together, and divided by the total number of female workers. This sum yields the proportional gain for women if a comparable worth policy is implemented. A similar calculation was made for male workers as well. The proportional gains were then used to determine the percentage gain in female earnings relative to male earnings once comparable worth is enacted. These calculations were completed for each of the four studies under review and are reported in the next section.

TABLE 1—THE IMPACT OF IMPLEMENTING COMPARABLE WORTH (CW) ON THE RELATIVE EARNINGS OF WOMEN AND MEN (Shown in Percent)

Studies	The Percent Gain For Women under CW[a]	Earnings Ratio Between Women and Men: Before CW[b]	Earnings Ratio Between Women and Men: After CW[c]	Cost of CW as a Percent of Payroll[d]
Iowa	12	74	82	6
Michigan	14	79	88	7
Minnesota	21	74	88	9
Washington	21	77	90	10
Average	17	76	87	8

Source: Author's calculations from the state's job evaluation studies, and the National Committee on Pay Equity's report (undated).

[a]The Percentage Gain for Women under CW equals $w \times 100$, were w equals: $\Sigma_i[(\hat{S}_{Li} - S_{Li})/S_{Li}] \times (n_{wi}/N_w)$, where i equals the set of all occupations; $\hat{S}_i$ equals the predicted occupational salary for i under comparable worth; S_i equals the current occupational salary for i; n_{wi} equals the number of female workers in i; N_w equals the total number of female workers.

[b]Earnings Ratio Between Women and Men Before CW equals a/b, where a equals average full-time female earnings, and b equals average full-time male earnings.

[c]Earnings Ratio Between Women and Men After CW equals c/d, where c equals $a \times w$, and d equals $b \times z$, where z equals $\Sigma_i[(\hat{S}_{Li} - S_{Li})/S_{Li}] \times (n_{mi}/N_m)$, where i, $\hat{S}_i$, S_i are defined as above; n_{mi} equals the number of male workers in i; N_m equals the total number of male workers.

[d]The Cost of CW as a Percentage of Payroll equals: $(w \times a \times N_w + z \times b \times N_m)/(a \times N_w + b \times N_m)$, where w, a, N_w, z, b, N_m, are defined above.

III. Findings from Four Job Evaluation Studies

The first column of Table 1 presents the percentage increase in female earnings from implementing comparable worth in each state. It ranges from 12 percent in Iowa to 21 percent in Washington; the average for the four states is 17 percent. Thus, this study finds that implementing comparable worth would increase female earnings by an average 17 percent, which represents a substantial improvement for women.[2]

[2]All of the statistical calculations and conclusions presented in this section are based upon the potential, initial impact of implementing comparable worth. Ultimate impact of this policy will depend upon the extent of its implementation throughout the economy and its effect on male and female employment.

The second and third columns of Table 1 report the earnings ratio between full-time female workers and full-time male workers before and after the implementation of a comparable worth policy. Before such implementation, the earnings ratio ranges from 74 percent in Minnesota to 79 percent in Michigan. The average earnings ratio for these jurisdictions is 76 percent, indicating that full-time female workers earn, on average, 76 percent as much as full-time male workers. After implementation, the earnings ratio increases to 87 percent, eliminating 46 percent of the pay disparity between women and men in these states. This represents a substantial decline in the male-female earnings gap.

Table 1 also reports the estimated cost of implementing comparable worth. Costs are reported as a percentage of existing payroll expenses, and range from 6 percent in Iowa to 10 percent in Washington. If comparable worth were implemented, the average percentage increase in payroll costs for these four states would be 8 percent.

IV. Conclusions

Using four public-sector job evaluation studies, comparable worth policies are shown to eliminate almost half (46 percent) of the earnings gap in these states. Based upon this evidence, implementing comparable worth nationwide would substantially reduce the earnings gap between women and men. Previous studies had indicated that a national comparable worth policy would have little effect on the male-female earnings gap. However, serious methodological problems vitiate these results.

REFERENCES

Ehrenberg, Ronald G., and Smith, Robert S., "Comparable Worth in the Public Sector," NBER, Working Paper No. 1471, September 1984.

Johnson, George, and Solon, Gary, "Pay Differences Between Women's and Men's Jobs: The Empirical Foundations of Comparable Worth Legislation," NBER Working Paper No. 1472, September 1984.

Steinberg, Ronnie J., "A Want of Harmony: Perspectives on Wage Discrimination and Comparable Worth," in Helen Remick, ed., *Comparable Worth and Wage Discrimination: Technical Possibilities and Political Realities*, Philadelphia: Temple University Press, 1984, 3–27.

National Committee on Pay Equity, *The Cost of Pay Equity in Public and Private Employment*, mimeo., undated.

Sex Differences in Urban Commuting Patterns

By MICHELLE J. WHITE*

This paper takes a new look, both theoretical and empirical, at the general question of what determines the pattern of urban workers' commuting journeys and at the specific question of how women workers' commuting journeys differ from those of men.

Commuting journey length for urban workers has proved difficult to model because it stands at the intersection of urban and labor economic theories concerning the spatial location patterns of jobs and housing. Urban economists view workers as having fixed job locations at the center of the city and being compensated for longer commuting journeys by lower housing prices in the suburbs. Labor economists, in contrast, tend to view workers as having fixed residential locations and being compensated for longer commuting journeys by higher wages at more distant jobs. See J. Madden and myself (1980) and Albert Rees and George Shultz (1970). The model presented here allows both locations to be determined simultaneously and both types of compensation for commuting to occur.

The problem gains an additional layer of complexity when sex differences in commuting patterns are considered, since sex differences in length of work trip are pronounced. Women workers have shorter commuting journeys, are more likely to take public transportation, and are more likely to work part time, therefore commuting at off-peak hours. Women workers also are more likely than men workers to have spouses who work and/or children at home; either of which may restrict their residential or job mobility. Also women workers earn less than men. This reduces their purchasing power in the housing market and therefore affects their job location and commuting possibilities. See Madden (1981).

I. Theory of Residential and Job Location Choice

Consider first an individual household's residential location decision. Households determine their residential location by maximizing a utility function, defined over housing, a composite good, and leisure time. The urban economics literature has shown that if 1) all households have one worker whose job is at the city center, and 2) all households have identical tastes and the same wage rate, then a market equilibrium housing price gradient exists which makes all households indifferent concerning both the distance and the direction from the city center at which they locate. This housing price gradient is denoted $p(u)$, where u is residential distance from the city center and p is the per unit price of housing. $p(u)$ has its maximum value exactly at the city center, it declines at a decreasing absolute rate with distance from the center, and it is identical in all directions.

Extending the model, if households still have identical tastes and job locations but there are several wage-skill levels, then the housing price gradient which makes households indifferent over all locations differs by income level. This causes households having different wage rates to occupy different locations. Each wage class prefers to locate over a range of distances having the shape of a ring around the center, rather than everywhere in the city. In general, richer households have flatter price gradients and occupy more distant rings.

However, if an individual household's tastes differ from those of households generally, then the market equilibrium housing price gradient will not make it indifferent over all or a range of residential locations. Instead it will prefer one or two particular locations over all others. For example, sup-

*Department of Economics, University of Michigan, Ann Arbor, MI 48109. I am grateful to the Alfred P. Sloan Foundation and the Urban Research Center, New York University, for research support and to Don Negri for research assistance.

pose a household has two workers rather than one, and both work at the city center. Then its commuting costs are higher than those of single-worker households having the same total income, so closer-in residential locations will be preferred. If it has two workers but one works in the suburbs, then it will prefer a residential location between the two jobs and probably nearer the suburban job, since housing prices are lower there.

Turn now to the pattern of workplace locations. Following the urban economics literature, I assume that there is some productive advantage to firms in being located at the city center. This could be because the center provides the broadest access to specialized services or skilled workers, or because it has the best transportation facilities. However, some employers have an incentive to move jobs to suburban locations. This is because in the suburbs they can pay lower wages to workers whose commuting trips are shortened. (Other prices also change for suburban firms, but we ignore them here.) The set of workers having the largest commuting cost reductions are those who live further from the center than the suburban job location and in the same direction away from the center. If the location chosen by the employer is v miles from the center, then each work day these workers save $2(m + w(v)/s)v$ in commuting costs; where m is monetary commuting expenses per mile each way, s is the speed of commuting per mile, assumed to be constant at all locations, $w(v)$ is the value of workers' time per hour, which will turn out to vary with job location, and v is the reduction in commuting distance each way that results from working at a suburban job.

Suppose some firms suburbanize, but they spread out in different directions and at different distances from the center. Each employs only workers who live further from the center than the firm and in the same direction away from the center. (Large firms are less likely to find suburban locations attractive than small firms, since by moving out they lose access to workers who live in other directions from the center.) Then there will be a market equilibrium spatial wage gradient, $w(v)$, determined by the process of location choice by employers and workers. The wage gradient has its maximum at the city center, declines with distance from the center and is uniform in all directions. It can be shown to decline at a decreasing rate with distance from the center. This is because from a fixed residential location, workers demand larger wage increments as they commute further inward towards the center. More time spent commuting reduces the total time available for leisure and work. Given diminishing marginal utility of leisure and goods consumption, greater loss of time must be compensated at higher and higher wage increments. (See my 1985 paper for a more detailed exposition of the model.)

Individual workers determine their job locations by maximizing their utility functions, taking as given the market wage gradient and the market housing price gradient (which has the same shape as in the centralized employment case). This means that, from a given residential location, they receive higher wages in return for commuting further, but only if they commute towards the center of the city. Out-commuting results in a lower wage per extra mile travelled, while circumferential commuting results in a constant wage regardless of miles travelled. If individual workers' tastes and other characteristics are the same as those of workers generally, then from fixed residential locations they will be indifferent among all jobs located between their residences and the city center. Extra commuting results in a higher wage just sufficient to offset the money cost and loss of leisure time of the extra distance travelled. From given job locations, workers face lower housing prices if they move their residences further out and commute greater distances, as long as they commute in an inward direction.

The model thus implies that workers whose households have typical tastes will be indifferent across all residential locations and across all job locations involving only in-commuting. Extending the model to include jobs involving multiple skill levels, each skill level will have a separate wage gradient. (However, firms hiring workers of different skill levels will be mixed at particular locations rather than segregated, as long as paying higher wages is not closely correlated with firms' willingness to pay for land.) Then

workers will be indifferent across all residential locations in the ring occupied by their income group and across all job locations involving only in-commuting from that ring. However, workers whose tastes are atypical will not be indifferent among residential and job locations. These workers will tend to prefer particular job or housing locations and particular commuting journey lengths. As an example, if the worker is a female head of household, then she may prefer a short commute because of heavy responsibilities at home. The market wage gradient may not be steep enough to induce her to commute more than the minimum distance.

The model's main conclusion concerning length of commuting trips is that the taste and demographic factors which differentiate individual households or workers from households or workers generally are the important explanatory variables determining commuting behavior. Only these prevent households and workers from being indifferent across a range of residential locations, across all job locations involving in-commuting, and therefore across a wide range of commuting journey lengths. To test this conclusion empirically, data are required for a sample of workers who are all located in the same metropolitan area, since the indifference property applies only across commuting journey lengths in a single city. I tested the model for several cities using data from the 1980 Annual Housing Survey. Due to space constraints, only the results for New York City are presented here.

The estimated equations explain length of commuting journey (in minutes), using income, taste, and demographic factors as explanatory variables. Because the model has no particular implications concerning functional form, the estimated equations are linear. The model is a reduced form. Neither the worker's wage nor the household's housing price is included as an explanatory variable. Since households maximize utility subject to exogenously determined market wage and housing price gradients, their actual wage and housing price variables each represent points chosen from the relevant schedules. These choices are therefore endogenous and including them in the equation would bias the results. As a result, the estimated coefficients in the commuting time regressions do not hold residential and job locations fixed. This means that if, for example, number of children increases, then the predicted change in the worker's commuting journey will incorporate the effect of changes in job or residential location that might be expected to occur as a result of the extra child, such as the worker's household moving to the suburbs.

In order to focus on sex differences in commuting behavior, separate equations are estimated for male and female workers. All workers in the data set are household heads, which biases the model against finding sex differences in commuting behavior since differences attributable to the behavior of secondary as opposed to primary workers are eliminated. The demographic variables are whether the household has a secondary worker or not (*P2WORKS*), whether there are preschool age children present or not (*YCHILD*), how many children under 18 are present (*NCHILD*), and a term interacting *P2WORKS* and *NCHILD* (*P2WCHILD*). Other variables are total family income (in thousands) in log form (*LINCOME*), whether the household head is black (*BLACK*) or is Spanish (*SPANISH*), whether the household owns its housing unit or rents (*OWNER*), and how many years since the household moved to its current housing unit (*YRSINHU*). No mode of travel variables are included, since choice of mode is also viewed as being endogenously determined by the same factors which explain commuting time.

The predicted effect on the household head's commuting journey of a secondary worker in the household could go in either direction, depending on where the second job is located. If both jobs are at the center, then the head's commuting journey is likely to be shorter. However if the head's job is at the center and the other job in the suburbs, then the effect of the second job may be to lengthen the head's journey if the household locates near the suburban job. A third possibility is that the second job has no effect on the head's commuting journey. More children

and especially young children are often thought to decrease women workers' commuting journeys, but this prediction is usually made for women who are secondary rather than primary workers. Higher income and owning housing are both expected to lengthen workers' commuting journeys, since both are associated with higher housing demand, which makes the suburbs' lower housing prices attractive. The variables for being black and Spanish are included since workers in each group are likely to locate in particular neighborhoods. But from these neighborhoods, most commuting may be outward or circumferential, making workers prefer to commute as little as possible. Of the other variables, longer residential tenure is likely to lengthen workers' commuting journeys if it implies less willingness to relocate when the worker changes jobs. It is included since New York has rent control, which holds down actual relative to market rent to a greater extent as households stay in the same apartment longer.

TABLE 1—COMMUTING JOURNEY LENGTHS FOR MALE AND FEMALE HOUSEHOLD HEADS, NEW YORK CITY, 1980[a]

	Females	Males
P2WORKS	.82	−.97
(.24, .45)	(1.84)	(1.05)
*NCHILD**	−1.00	2.73
(.56, .90)	(.87)	(.55)
*YCHILD**	8.53	−.56
(.08, .20)	(2.47)	(1.16)
P2WCHILD	−1.78	−1.79
(.16, .39)	(1.55)	(.75)
*LINCOME**	1.36	4.44
(2.6, 3.2)	(.84)	(.65)
*BLACK**	9.13	5.35
(.27, .11)	(1.43)	(1.26)
*SPANISH**	6.81	.76
(.08, .07)	(2.26)	(1.52)
*OWNER**	−1.69	5.34
(.34, .66)	(1.52)	(.96)
YRSINHU	−.31	−.30
(6.4, 7.2)	(.18)	(.12)
Intercept	26.94	20.37
	(2.55)	(2.15)
R^2	.03	.06
N	1448	5291
dep. mean	30.91	37.66
SSE	812,190	4,044,048

[a]Mean values of independent variables for females and males, respectively, are shown in parentheses below each variable.

II. Estimation Results

Regression results are given in Table 1, where standard errors are in parentheses. Asterisks give results of a separate statistical test for whether the male and female coefficients of each variable are significantly different. Despite the fact that both samples consist entirely of household heads, the results show substantial differences in commuting patterns by sex.

Turning to the household composition variables, children and secondary workers affect the commuting journeys of male and female household heads differently. Each extra child increases the commuting journey of male heads by 2.7 minutes if there is no secondary worker in the household, but by only .9(= 2.7 − 1.8) minute if there is a secondary worker. Thus male-headed households tend to suburbanize as they have more children, but the effect is much smaller if someone else in the household works. If the household has no children, then the second job has no significant effect on the head's commuting journey. However this result does not necessarily imply that working wives are forced to find a job from the fixed household residential location. Since the presence of a secondary worker can either raise or lower the commuting journey length of the household head, these two effects may offset each other in the data set.

In contrast, female heads' commuting journey length is not significantly affected either by the number of children, by the presence of a secondary worker, or by both at once. But the presence of young children has a large and significant effect which is positive rather than negative as expected—young children increase the commuting journey of female household heads by 8 minutes or 26 percent.

Other results are that male workers who own housing have longer commuting journeys than male renters, but the commuting journeys of female owners are not significantly different from those of female renters.

Also higher income by itself has only a small effect on commuting journey length for male workers and a small and insignificant effect for female workers. The small effect of income is not surprising since firms employing higher income workers also have a stronger incentive to move out as workers' wages rise. The length of tenure variable shows, contrary to expectations, that longer tenure in the same housing unit is associated with a shorter commuting journey, by .3 minutes per extra year of residence for both sexes. Rather than being forced to commute further because of their unwillingness to move when they shift jobs, workers having long residential tenure appear to adjust by finding jobs near their homes, perhaps at a sacrifice in income. For both sexes, being black is associated with a large increase in commuting journey length.

Thus male and female workers' commuting patterns are quite different generally and show different patterens of responsiveness to the presence of children and secondary workers, even when workers of both sexes are household heads. Male workers' commuting journey length is significantly shortened by the presence of a second worker, but only if there are children in the household. Female workers' commuting journeys are unresponsive to any of the demographic variables except for the presence of young children.

REFERENCES

Madden, J., "Why Women Work Closer to Home," *Urban Studies*, 1981, *18*, 181–94.

______ **and White, M.,** "Spatial Implications of Increases in the Female Labor Force: A Theoretical and Empirical Synthesis," *Land Economics*, November 1980, *56*, 432–46.

Rees, A., and Shultz, G., *Workers and Wages in an Urban Labor Market*, Chicago: University of Chicago Press, 1970.

White, M., "Commuting Behavior in Cities with Decentralized Employment," working paper, Department of Economics, University of Michigan, 1985.

______, "A Model of Residential Location Choice and Commuting by Men and Women Workers," *Journal of Regional Science*, April 1977, *17*, 41–52.

Employment and Wage Effects of Involuntary Job Separation: Male-Female Differences

By NAN L. MAXWELL AND RONALD J. D'AMICO*

Recent compositional changes in industrial and occupational structures have created technological displacement for many workers. Accompanying these institutional changes is increased foreign competition which has forced previously profitable firms to collapse. What happens to workers when competition and displacement create job loss? Many studies have found that with job termination, workers face long spells of unemployment and reentry into occupations with lower wages and fringe benefits; however, few studies have examined gender differences in the consequences of job termination. In fact, most research in this area has been case studies focusing solely on men. This study analyzes gender differences in employment and wages upon job termination. It answers the question, "Do women fare better or worse than men upon job termination?"

I. Framework

Within the human capital model of labor markets, the role of training is critical for understanding wage loss upon displacement. If an individual with a large portion of general human capital skills is displaced, the consequences of job termination will be less severe than for the worker who accumulated firm-specific skills. Because firm-specific skills do not increase the worker's productivity outside a particular firm, the worker with a large stock of firm-specific skill cannot transfer to a firm in which productivity (and hence wage) will be as high. On the other hand, a worker with a large stock of general training will be able to transfer productivity to another firm and will be less likely to suffer wage loss.

It is often argued that women's intermittent labor force participation causes investment in general, rather than firm-specific human capital. The notion is that such an investment pattern facilitates labor force movement without large wage loss. On the other hand, men, with continual labor force attachment, invest more heavily in firm-specific skills. Although this restricts mobility, the investment increases their wage in the particular firm at which they are employed. Thus, the human capital model suggests that women investing in fewer firm-specific skills than men will suffer less wage loss upon displacement because of the differential weighting of human capital investments—not because of gender.

A more institutional view argues that occupation and industry of displacement, not individual characteristics, determine wages and displacement consequences, and that discrimination plays a powerful role within these institutions. Within this view, occupational internal labor markets form which increase wages yet restrict interfirm mobility at comparable wages. Unfortunately, women have difficulty obtaining these high-wage positions. Similar institutions arise along industrial lines with increased wages available for workers in larger firms. Furthermore, for women, differential affirmative action enforcement may result in economic rents in larger, monitered firms. While women within internal labor markets and larger establish-

*Department of Economics, California State University, Hayward, CA 94542, and SRI International, Menlo Park, CA, respectively. We thank Jeff Golon and Michael Motto for their outstanding research and programming assistance. The majority of this project was funded by Department of Labor under the Comprehensive Employment and Training Act while D'Amico was located at the Center for Human Resource Research, Ohio State University; the remainder of the project was funded by Denison University while Maxwell was located there. We bear sole responsibility for viewpoints of the document.

ments benefit from increased wages, gender discrimination prohibits most women from employment in these sectors. Thus, women displaced from these labor market institutions face double jeopardy. First, unlike men, they face gender discrimination when obtaining an equivalent position. Second, like men, they are employed in occupations and industries in which predisplacement wages cannot be obtained elsewhere.

Human capital and labor market institutions, through their effect on reservation wage, also play an important role in determining unemployment after displacement. When a worker's reservation wage is greater than market wage, unemployment will result. As unemployment continues, the reservation wage declines until it is equivalent to market wage and employment occurs. However, the declining reservation wage may approach the shadow wage in home production before reaching market wage and labor force participation will cease. Since women often specialize in home production while most men specialize in market work, a woman's postdisplacement market wage is more likely to be below her home productivity; hence, increased unemployment is more likely to result in labor force withdrawal for women than men.

Irrespective of gender, human capital and labor market institutions help set the reservation wage and, hence, determine unemployment following displacement. If a worker has only general human capital skills, the predisplacement wage reflects labor market worth and duration of unemployment will be relatively short. However, if skills are heavily weighted with firm-specific human capital, or if displacement occurs from an occupation within an internal labor market or an industry paying economic rents, the reservation wage will exceed market worth and duration of unemployment will be longer.

This study examines the role human capital and institutional factors play in explaining the consequences following involuntary job termination. To the extent that human capital and institutional factors do not explain these consequences, socioeconomic factors associated with gender are operating. These other factors can range from labor market discrimination to gender-differing job-search techniques to societal and individual attitudes toward women's roles. We consider these factors as "gender differences" and study human capital, institutional, and gender effects on the consequences of displacement.

II. Data and Methodology

Unfortunately, it is difficult to quantify firm-specific human capital, labor market institutions, and home production. In this study we use industry and occupation variables as proxies for institutional characteristics, tenure as a proxy for firm-specific human capital, and an individual's predisplacement relative wage in the household (i.e., wage and salary income/total family income) as a proxy for relative specialization in the labor market. To the extent that nongeneral human capital accumulation raised predisplacement wages, we insert predisplacement hourly rate of pay in the multivariate estimation equations.

Data are taken from the young men and young women's panels of the *National Longitudinal Surveys (NLS)*. These nationally representative surveys collected data over a 15-year period from 5000 men aged 14–24 in 1966 and 5000 women aged 14–24 in 1968. The longitudinal nature of the data enabled construction of a sample pooled across the survey period. To be included in this sample, a worker must be involuntarily displaced during the survey period.

Displacement is conceptualized as occurring to a "mainstream" worker with a relatively steady employment relationship involuntarily severed, and for whom the prospect of being rehired by the same employer is close to nil. Operationally, the following criteria were used: 1) the worker was at least 20-years-old and not enrolled in school in the survey before the employment relationship was broken, 2) the worker was employed at least 24 months with the same employer, 3) the employment ended with the reason given as being involuntary, and 4) the worker did not return to this employer in any subsequent survey year. To prevent the possibility of error term correlation associated with pooled cross-section analysis, we

allowed a displaced individual to enter the sample only once with the final displacement counting in the analysis. (Only 1 woman and 3 men had multiple displacements.)

For displaced respondents, we estimated a predicted wage in the first $(t+1)$ and third $(t+3)$ years after displacement on the basis of predisplacement wages, age, education, race, SMSA residence, industry of employment, and hours worked. These prediction equations were estimated on the basis of coefficients generated from the sample of nondisplaced respondents who met the age, nonenrollment, and tenure restrictions used for the displaced sample. Since wage-growth profiles vary according to business cycle conditions and gender, all prediction equations were estimated separately for nondisplaced males and females in each survey year. Coefficients for prediction equations for displaced workers were obtained from the wage equation estimated during the period the displacement occurred. The difference between predicted and actual wages in the postdisplacement periods was interpreted as the wage loss due to displacement.

Because employment data are not continuous throughout the survey period, employment status was operationalized as survey-week activity at the first survey and approximately 3 years after displacement. Thus, we know if the displaced worker was employed, unemployed, or out of the labor force approximately 1 year and approximately 3 years after displacement.

III. Results

Tabular results showed that, like the labor market as a whole, gender differences in our displaced sample exist. A higher percent of displaced men are professional/managers or craft workers and come from the construction sector, while a higher percent of women are clerical/sales and come from the service sector. Displaced men have longer tenure and their predisplacement wage and salary income represents a larger percent of total family income than women.

While our sample is restrictive, evidence suggests that gender differences exist in the risk of displacement with females having a much lower probability of being displaced than males (2.54 percent of the females and 3.99 percent of the males were displaced). With the exception of the agricultural industry and sales and farm occupations, females had lower displacement rates in all industries and occupations.

Since women often have less tenure and firm-specific human capital than men, their lower displacement rates are somewhat surprising. Usually, employers are less willing to lay off workers with firm-specific human capital and layoffs often are seniority based, hence, we expect women to have higher displacement rates than men. We surmise that, because we are analyzing displacement of the "mainstream" worker, our women at risk for displacement are more serious labor market participants than the "typical" woman worker. This being the case, our low female displacement rates may not be reflective of the norm, but of our labor force committed sample.

Since these women appear to be committed to the labor market, gender differences in employment after job termination are startling (Table 1). By the first survey after displacement, women are 3 times as likely as men to be unemployed. While part of this difference may be due to the fact that the measurement period after displacement is shorter for women than men, the differential is still $2\frac{1}{2}$ times greater by the third survey after displacement. In fact, while female unemployment rates drop dramatically over the 2-year period (from 85.27 to 21.66 percent), their unemployment rate 33 months after displacement is only slightly below male unemployment 8 months after displacement (21.66 and 26.97, respectively).

Given high female unemployment rates, it is not surprising that displaced women are more likely than men to leave the labor force. By the first survey after displacement, about 40 percent of the women and 3.8 percent of the men have left the labor force. Because of high unemployment rates and low labor force participation rates, 5 months after job termination only 8.69 percent of the women are reemployed. While this figure increases to 55.29 percent by 33 months after displacement, it never approaches male em-

TABLE 1—EMPLOYMENT STATUS OF DISPLACED WORKERS AT SURVEYS $t+1$ AND $t+3$[a]

Employment Status	Males		Females	
	$t+1$	$t+3$	$t+1$	$t+3$
Employed	70.3	87.8	8.7	55.3
Unemployed (as a percent of all respondents)	25.9	9.0	51.1	15.3
Out of Labor Force	3.8	4.3	40.2	29.4
Unemployment Rate	26.97	9.25	85.27	21.66
Mean Months Elapsed from Stop Date to Survey	8.00	34.70	5.15	32.66

[a]Shown in percent—percent is computed on the number of interviewed respondents only.

ployment of 70.26 percent 8 months after displacement.

In order to test whether these differential displacement consequences are due to human capital, institutions, or gender, a multinominal logit model of employment status (employed, unemployed, out of the labor force) and an ordinary least squares model of wage loss (predicted wage minus actual wage) were estimated. Because so few women were reemployed after displacement, sample size did not permit gender stratification of equations. In order to control for wage biases created by out of the labor force or unemployed respondents, two sample selection bias terms were inserted into the wage loss equation.

Results of the multinominal logit showed that net of human capital (tenure, education, labor market specialization, predisplacement wage, age, and race) and institutional influences (industry, occupation, industry growth, and changing residence), females have a greater difficulty obtaining employment in the short run than do males. However, about 3 years after displacement, these gender differences are no longer significant. By $t+3$, human capital influences predict employment status with tenure, education, and predisplacement wage negatively associated with the probability of being employed. Likewise, individuals with a large portion of the family income determined by their wage and salary income are more likely to be employed in $t+3$. On the other hand, workers previously employed as clerical/sales are more likely to be unemployed or out of the labor force as time passes.

Thus, while being female per se does not influence employment status 3 years after displacement, characteristics associated with female workers are influences. Typically females are employed as clerical/sales, have low rates of pay and tenure, and contribute relatively little to family income. Since these characteristics are all associated with unemployment and out of the labor force status 3 years after displacement, the typical female will not fare well if displaced from her job. In fact, these results may be *understating* gender differences in postdisplacement unemployment since our sample of females appears more committed to the labor force than the "average" female worker.

Because so few women are reemployed after displacement, there is very little variance in the gender variable in the wage loss equation making it extremely difficult to gain significance in the model. However, the sign of the variable indicates that females have greater wage loss than males, controlling for human capital (tenure, education, age, and race) and institutional (changing industry, occupation, and residence) differences, and the variable approaches conventional significance levels by the third survey after displacement ($t=1.78$, $p \leq .10$). Thus, there is some indication that being female increases wage loss after displacement.

IV. Summary

Results of this study show that, while males may have increased displacement rates, once females lose their jobs they are more likely to have difficulty recovering their initial labor market positions. Striking employment differentials between the sexes exist after displacement with female unemployment rates about $2\frac{1}{2}$ times larger than males. With prolonged unemployment and relatively high levels of home productivity, women are much more likely to drop out of the labor force

than men. Thus, their employment rates 33 months after displacement do not approach male employment rates 8 months after displacement. Much of this differential can be attributed directly to gender or to gender-related characteristics. While the evidence of gender differences in wage loss is weaker, we provide evidence that, net of human capital and institutional influences, displaced females also suffer greater wage loss than males.

Thus, the answer to the question, "Do women fare better or worse than men upon job termination?," is that they fare considerably worse. Policymakers need to be aware of this when analyzing policies using displacement studies based solely on men.

Generational Differences in Female Occupational Attainment—Have the 1970's Changed Women's Opportunities?

By NADJA ZALOKAR*

The issue of equal pay for comparable work has in recent years come to the forefront of debate in the political arena. Crucial to estimating the potential consequences of comparable worth legislation is an understanding of the causes of sex differences in occupations.

Earlier studies have found evidence that sex differences in labor force attachment may explain sex differences in occupations. (See Solomon Polachek, 1977, 1979, 1981; and my 1984 paper.) However, Paula England (1982) and Mary Corcoran et al. (1983) find that women with high labor force attachment are no more likely than other women to be in male occupations. This result suggests that when choosing occupations, women may face constraints in the form either of direct labor market discrimination preventing them from entering male occupations, or of a socialization process through which women and men acquire different tastes for occupations.

Virtually all the evidence discussed above is derived from data on women who came of age in the late 1940's and early to mid-1950's. This paper seeks to determine whether young women today continue to face the same constraints. To do so, the structure of occupational attainment for two cohorts of women is compared. The older cohort, 2894 women taken from the *National Longitudinal Survey of Mature Women*, was born in the decade before World War II (1929–1937), and the younger cohort, 2632 women taken from the *National Longitudinal Survey of Young Women*, was born in the decade following (1944–52). The paper compares the two groups when each cohort was aged 30–38: in 1967 for the older cohort and in 1982 for the younger cohort. Thus, the observations on the two groups are separated by the crucial decade of the 1970's when the effects of equal employment opportunity legislation and the women's movement are thought to have taken hold.

Section I compares the two cohorts and finds that the younger women are more attached to the labor force and enter more skilled, "less female" occupations than their older counterparts. Section II attempts to determine the underlying reasons for the younger women's increased occupational attainment.

I. Generational Differences in Labor Force Attachment and Occupational Attainment

The younger women are much more attached to the labor force than their older counterparts. At ages 30–38, they have on average worked $1\frac{1}{2}$ years more (9.53 compared to 8.11 years), acquired $1\frac{1}{2}$ years more education (12.74 compared to 11.11 years) and spent $3\frac{1}{2}$ fewer years at home (5.59 compared to 9.03) than women in the older cohort. Of the younger cohort, 67.5 percent and of the older cohort, 48.5 percent are employed at ages 30–38. One-third of the younger cohort has spent less than 2 years at home, compared to one-fifth (19.4 percent) of the older cohort. Under 20 percent (19.3 percent) of the younger cohort has spent more than 10 years at home, compared to over 40 percent (42.2 percent) of the older cohort.

Two concepts of occupational attainment are used to compare the cohorts. First, the proportions female of their 3-digit Census occupations are compared to determine whether the younger cohort has entered more "male" occupations. The older women's first occupations are on average 67.6 percent

*Department of Economics, University of Florida, Gainesville, FL 32611.

female and their occupations at ages 30–38 (1967 occupations), 66.9 percent female. The younger women's first occupations are 68.8 percent female, about the same as the older women's. However, the younger women move into more male occupations later on: at ages 30–38 (in 1982), their occupations are on average 60.9 percent female.

Second, the human capital or skill requirements of the women's occupations are compared to determine whether the younger cohort has entered more skilled occupations. Two dimensions of human capital or skill requirements are considered. The first is "general" human capital, or skills that are useful in a wide variety of occupations. The second is "occupation-specific" (hereafter, "specific") human capital, or skills that are useful only in one occupation or in a small group of closely related occupations. The older cohort's first occupations require an average of 11.28 years of general human capital and .94 years of specific human capital investment. The skill levels of their occupations at ages 30–38 are virtually unchanged: 11.41 years of general and 1.01 years of specific human capital. The younger women enter more skilled occupations than their older counterparts. Their first occupations require 12.01 years of general and 1.16 years of specific human capital, and their occupations at ages 30–38 require 12.23 years of general and 1.53 years of specific human capital. Moreover, the younger women move into more skilled occupations as they age: there is a 32 percent increase in specific human capital levels between their first occupations and their occupations at ages 30–38.

In sum, younger women exhibit much greater labor force attachment and occupational attainment than the older women. While the increase in the skill levels of their occupations is striking, the decrease in femaleness of their occupations is small.

TABLE 1—MEAN OCCUPATIONAL ATTAINMENT BY YEARS OF HOME TIME

Years of Home Time[a]	Percent of Sample	Percent Female	General Human Capital[b]	Specific Human Capital[c]
NLS Mature Women				
< 2	19.4	67.2	11.88	1.46
2–5	12.2	68.9	11.51	1.10
5–7	10.1	69.0	11.58	1.08
7–10	16.1	67.7	11.51	.97
> 10	42.2	65.2	11.03	.74
NLS Young Women				
< 2	33.3	56.0	13.06	2.15
2–5	21.1	61.1	12.30	1.59
5–7	11.1	59.7	11.81	1.24
7–10	15.1	64.0	11.76	1.13
> 10	19.3	67.2	11.34	.89

[a]Years spent at home by ages 30–38 (not in school or in the labor force).

[b]Median number of years of schooling in 1960 of all workers in the 3-digit Census occupation held at ages 30–38.

[c]Number of years of training corresponding to the Specific Vocational Preparation (SVP) category of the 3-digit Census occupation held at ages 30–38.

II. Sources of Women's Increased Occupational Attainment

Section I finds that by ages 30–38 women in the younger cohort entered both more skilled and more male occupations than women in the older cohort. This section attempts to isolate the underlying reasons for their increased occupational attainment.

According to a simple human capital model in which women choose human capital investment levels and years of home time to maximize a utility function over lifetime income and home time, women's occupational attainment can change over time for one of the following three reasons. First, there can be an exogenous increase in the proportion of women with high relative preferences for lifetime income (low relative preferences for home time). More women will take smaller amounts of home time, and their human capital investments (occupational attainment) will increase accordingly. Second, there can be an exogenous change in the wage function facing women, creating incentives for women to both decrease their home time and increase the skill levels of their occupations. Third, there can be a decrease in the costs of entering skilled occupations (or, alternatively, a change in women's tastes causing women to act as though the costs of

entering skilled occupations had decreased), which also creates incentives to decrease home time and increase occupational attainment.

Table 1 allows consideration of the first possible explanation for the observed increase in occupational attainment—there was an exogenous increase in the proportion of women with low relative preferences for home time. Table 1 shows for both cohorts the mean values of the percent female and the general and specific human capital levels of the occupations entered by women with different amounts of home time. For both cohorts, women who take less home time enter occupations requiring more human capital, particularly specific human capital, consistent with human capital theory's predictions. For the older cohort, there is no consistent relationship between home time and femaleness of occupations, consistent with the evidence cited above. However, for the younger cohort, women who take less home time also enter less female occupations.

If an exogenous change in relative preferences for home time accounts for the changes in occupational attainment between the two cohorts, then the mean values of the occupational attainment measures should be the same for women with the same amount of home time from either cohort. However, Table 1 shows that women in the younger cohort enter less female, more skilled occupations than their older counterparts in the same home time category. Thus, an exogenous change in relative preferences for home time cannot fully account for the differences in the occupational attainment of the two cohorts.

Table 2 shows estimates of wage equations for the older cohort in 1967 and the younger cohort in 1982, permitting examination of the second possible explanation for the younger cohort's increased occupational attainment. There are several differences between the two wage equations. An extra year of work experience increases the 1982 wage by 4.6 percent, more than the comparable figure of 1.5 percent for 1967, and the returns to specific human capital are slightly larger, creating incentives for the younger cohort to take less home time and enter more skilled occupations than the older cohort. Yet, the returns to general human capital fell between 1967 and 1982, creating the opposite incentives. The wage penalty for being in a female occupation also fell.

For both wage equations, a woman with the average home time and occupational attainment of the younger cohort would earn approximately 50 percent more up to age 35 than a woman with the average home time and occupational attainment of the older cohort. Thus, while there exists a substantial economic incentive for women to choose less home time and greater occupational attainment, this incentive does not appear to have changed over time. Although this result depends on a fairly crude simulation of lifetime earnings, it suggests that the changes in the wage equation between 1967 and 1982 probably had little effect on women's choices of home time and occupational attainment.

TABLE 2—WAGE EQUATIONS[a]

	NLS Mature Women	*NLS Young Women*
Intercept	−.1487	5.7709
	(−1.40)	(69.17)
Experience[b]	−.0023	.0231
	(−.70)	(6.34)
Home Time[c]	−.0177	−.0235
	(−5.38)	(−5.86)
General Human Capital	.0780	.0301
	(12.01)	(5.40)
Specific Human Capital	.0314	.0461
	(3.29)	(6.40)
Labor Force[d]	.2093	.1728
	(9.35)	(8.88)
Black[e]	−.1672	−.1161
	(−6.63)	(−5.21)
Percent Female of Occupation	−.2237	−.1292
	(−5.54)	(−4.10)
R^2	.3767	.2258
F	105.69	125.67
Number in Sample	1231	1747

[a]Ordinary least squares estimation, dependent variable is the natural logarithm of the 1967 wage for the *NLS Mature Women* and the 1982 wage for the *NLS Young Women*. The *t*-statistics are shown in parentheses.

[b]Years of labor force experience by ages 30–38.

[c]Years of home time by ages 30–38.

[d]Dummy variable equal to 1 if lived in an area where with a labor force of more than 200,000.

[e]Dummy variable equal to 1 if woman is black.

In sum, the younger cohort's increased occupational attainment is probably not due to an exogenous change in women's tastes for home time or to a change in the wage function facing women—instead, it seems likely that their increased occupational attainment is due to the third possibility, a decrease in women's costs of entering (or an increase in women's tastes for) more skilled, less female occupations.

III. Conclusion

The primary source of women's increased occupational attainment during the 1970's was a decrease in women's costs of entering (increase in women's tastes for) more skilled, less female occupations. Moreover, unlike their older counterparts, younger women who take less time out of the labor force enter more male occupations. Thus, it appears that women experienced a lessening of the "constraints" on their choice of occupations during the decade of the 1970's. However, the results of this paper do not allow us to determine whether the change was one of "costs" or one of "tastes." Either decreased occupational discrimination against women (costs) or increased occupational ambitions (tastes) could be responsible for the increase in occupational attainment.

The 1970's were a decade of improvement in women's occupational opportunities, suggesting that the case for comparable worth legislation has weakened over time. However, the occupational distributions of men and women continue to be very different, even when their differences in labor force attachment are controlled for. Before any ultimate conclusions about the need for comparable worth legislation can be drawn, further research aimed at sorting out the roles of tastes and discrimination in determining sex differences in occupations is necessary.

REFERENCES

Corcoran, Mary, Duncan, Greg J. and Ponza, Michael, "Work Experience and Wage Growth of Women Workers," in Greg J. Duncan and James N. Morgan, eds., *Five Thousand American Families — Patterns of Economic Progress*, Vol. 10, Ann Arbor: ISR, University of Michigan, 1983.

England, Paula, "The Failure of Human Capital Theory to Explain Occupational Sex Segregation," *Journal of Human Resources*, Summer 1982, *17*, 358–70.

Polachek, Solomon, "Occupational Segregation Among Women: A Human Capital Approach, Paper No. 77–4, University of North Carolina, 1977.

______, "Occupational Segregation Among Women: Theory, Evidence and a Prognosis," in Lloyd, Cynthia B. et al., eds., *Women in the Labor Market*, New York: Columbia University Press, 1979, 137–57.

______, "Occupational Self-Selection: A Human Capital Approach to Sex Differences in Occupational Structure," *Review of Economics and Statistics*, February 1981, *63*, 60–69.

Zalokar, Nadja, "Male-Female Differences in Occupational Choice and the Demand for General and Occupation-Specific Human Capital," mimeo., June, 1984.

The Existence of Equilibrium with Price-Setting Firms

By ERIC MASKIN*

Ever since Joseph Bertrand (1883), economists have been interested in static models of oligopoly where firms set prices. Francis Edgeworth's 1925 critique of Bertrand recognized, however, that, except in the case of constant marginal costs, there are serious equilibrium existence problems when firms produce a homogeneous good. In particular, Edgeworth proposed a modification of Bertrand's model in which firms have zero marginal cost up to some fixed capacity. He showed that, unless demand is highly elastic, price equilibrium may fail to exist.

Mixed strategies provide one way of avoiding this nonexistence problem, as various authors have noted. Martin Beckmann (1965), for instance, explicitly calculated mixed strategy equilibria in a symmetric example of the Bertrand-Edgeworth model. However, a general treatment of mixed strategies has suffered from the fact that the standard equilibrium existence lemmas (see, for example, K. Fan, 1952, and I. Glicksberg, 1952), require continuous payoff functions, whereas, in price-setting oligopoly, payoffs are inherently discontinuous—the firm charging the lowest price captures the whole market.

Recently, Partha Dasgupta and I (1986) and Leo Simon (1984) developed several existence theorems for discontinuous games. Dasgupta and I used one of the theorems to establish the general existence of mixed strategy equilibrium in the Bertrand-Edgeworth model when market demand as a function of price is continuous, downward sloping, and equal to zero for a sufficiently high price. In their study of Bertrand-Edgeworth competition in large economies, Beth Allen and Martin Hellwig (1983) extended this result to demand curves that do not necessarily intersect the horizontal axis and need not slope downward.

There have been several treatments of cost functions more general than the Bertrand-Edgeworth variety. R. Gertner (1985) established the existence of symmetric equilibrium in a model where firms are identical, have convex or concave costs, and choose output levels at the same time as prices. Also in a model of identical firms, H. Dixon (1984) proved existence when firms have convex costs and produce to order, that is, produce *after* other firms' prices are realized.

In this paper, I present some existence results that do not require symmetry and permit fairly general cost functions. These findings pertain both to the simultaneous choice of price and production level and to the formulation where a firm's output is set only after it knows others' prices. Existence in the former case is proved by direct application of the Dasgupta-Maskin/Simon theorems, as are the existence results mentioned above. The latter case, however, requires some additional argument. I give a sketch of this argument below; the details and more general results can be found in Dixon and myself (1986).

I. The Model

For simplicity, I shall consider only two firms; all results generalize immediately to any finite number. Firms produce the same good, and firm i, $i=1,2$, has total cost function $c_i(x_i)$, where x_i is the firm's output

†*Discussant*: Richard Schmalensee, Massachusetts Institute of Technology.

*Department of Economics, Harvard University, Cambridge, MA 02138. I thank the Sloan Foundation and the NSF for research support. I am indebted to Martin Hellwig for helpful comments on an earlier version of this paper.

level. We assume

(1) $c_i(x)$ is continuous and nondecreasing; $c_i(0)=0$.

Firms face an industry demand curve $F: R_+ \to R_+$, where

(2) F is continuous with $F(0)=K$ for some $K>0$; $pF(p)-c_i(F(p))$ is maximized at $\bar{p}_i>0$, and, if there are multiple maximizers, $\bar{p}_i$ is the largest. Let $\bar{p}=\max p_i$.

Firm i may have a constraint K_i on its production level. In view of (2), we may assume, without loss of generality, that $K_i \le K$.

Firm i's strategy consists of choosing a price $p_i \in [0, \bar{p}]$ and supply $s_i \in [0, K_i]$. Given the firms' strategies the demand facing firm i is

(3) $$d_i(p_1, s_1, p_2, s_2) = \begin{cases} F(p_i), & \text{if } p_i < p_j \\ G_i(p, s_1, s_2), & \text{if } p_1 = p_2 = p \\ H_i(p_1, p_2, s_j), & \text{if } p_i > p_j, \end{cases}$$

where G_i is a function such that

(4) $G_1(p, s_1, s_2)+G_2(p, s_1, s_2)=F(p)$; if $s_i>0, G_i>0$; if $s_i \ge s_j, G_i \ge G_j$; and $\min\{G_i(p, s_1, s_2), s_i\}$ is continuous in s_i,

and $H_i(p_1, p_2, s_j)$ is a function such that

(5) $H_i(p_1, p_2, s_j) \le F(p_i)$; $H_i(p, p, s_j) = F(p)-s_j$; and H_i is continuous.

The first line of (3) simply posits that the firm charging the lower price attracts the entire market demand.

The second line stipulates that, if firms charge the same demand, they split the market in some way that depends on their supplies. Condition (4) tells us that a firm's share is a nondecreasing function of its supply and that a positive supply implies a positive share. Two examples of G_i's satisfying (4) are

$$G_i(p, s_1, s_2) = \begin{cases} \dfrac{s_i}{s_1+s_2}F(p), & \text{if } s_1+s_2>0 \\ \dfrac{1}{2}F(p), & \text{if } s_1+s_2=0 \end{cases}$$

and $G_i(p, s_1, s_2) = \alpha_i F(p)$, where $\alpha_1+\alpha_2=1$.

The third line of (3) requires that the firm charging the higher price get less than full market demand. Moreover, if the two firms are charging approximately the same price, the demand facing the high-price firm is approximately full market demand at that price minus the supply of the other firm. Two examples of H_i's satisfying (5) are $H_i(p_1, p_2, s_j)=\max\{F(p_i)-s_j, 0\}$ and

$$H_i(p_1, p_2, s_j) = \max\left\{\frac{F(p_j)-s_j}{F(p_j)}, 0\right\}F(p_i).$$

The former rule is called parallel rationing (see Richard Levitan and Martin Shubik, 1972), whereas the latter is proportional rationing (see Edgeworth).

In the case where a firm sets production at the same time as price (production in advance), firm i's payoff is

(6) $$p_i x_i - c_i(s_i),$$

where

(7) $$x_i = \min\{s_i, d_i(p_1, s_1, p_2, s_2)\},$$

whereas in the case where it produces to order, the firm's payoff is

(8) $$p_i x_i - c_i(x_i).$$

II. Equilibrium in Discontinuous Games

Let us present a special case of the main existence theorems in Dasgupta's and my

article, and in Simon. For $i=1,2$, let A_i be a convex, compact subset of R^2. The set A_i is player i's strategy space. Let $U_i: A_1 \times A_2 \to R$, the payoff function for player i, be (a) bounded and (b) continuous except at points $(p_1, s_1, p_2, s_2) \in A_1 \times A_2$ where $p_1 = p_2$. Assume, furthermore, that U_i is weakly lower semicontinuous in (p_i, s_i). That is, for any (p_1, s_1) there exists a sequence $\{(p_1^n, s_1^n)\}$ converging to (p_1, s_1) such that no two p_1^n's and no two s_1^n's are the same and such that, for any (p_2, s_2),

$$\text{(9)} \qquad \lim_{(p_1^n, s_1^n) \to (p_1, s_1)} U_1(p_1^n, s_1^n, p_2, s_2) \geq U_1(p_1, s_1, p_2, s_2),$$

and similarly for player 2. Finally, suppose that $\Sigma_{i=1}^2 U_i$ is upper semicontinuous.

PROPOSITION: *Given the stated assumptions, a mixed strategy equilibrium exists. That is, there exist probability measures* (μ_1^*, μ_2^*) *such that*

$$\int U_1(p_1, s_1, p_2, s_2)\, d\mu_1^*(p_1, s_1) \times d\mu_2^*(p_2, s_2) \geq \int U_1(p_1, s_1, p_2, s_2)\, d\mu_1(p_1, s_1) \times d\mu_2^*(p_2, s_2) \qquad \textit{for all } \mu_1 \textit{ on } A_1;$$

$$\int U_2(p_1, s_1, p_2, s_2)\, d\mu_1^* \times d\mu_2^* \geq \int U_2(p_1, s_1, p_2, s_2)\, d\mu_1^* \times d\mu_2, \qquad \textit{for all } \mu_2 \textit{ on } A_2.$$

III. Equilibrium in Oligopoly

Let us take $A_i = [0, \bar{p}] \times [0, K_i]$ in our oligopoly model. In the case of production in advance, we can readily verify that firms' payoffs satisfy the hypotheses of the above proposition. From (2), profits are clearly bounded. Because F, the H_i's and the c_i's are continuous, profits are continuous except where $p_1 = p_2$. If $p_1 = 0$ or $s_1 = 0$, then U_1 is continuous at (p, s_1, p_2, s_2), and so (9) holds automatically. If p_1 and s_1 are both positive, then $d_1(p_1, s_1, p_2, s_2)$ is lower semicontinuous from the left. That is, at points of discontinuity (where $p_1 = p_2$), firm 1's demand jumps *downward* for a sequence $\{p_1^n\}$ of prices converging to p_1 from below. Hence, (9) holds for any convergent sequence $\{(p_1^n, s_1^n)\}$ where p_1^n converges to p_1 from below. Finally, observe that discontinuities simply entail a shift in demand from one firm to the other. Thus, although U_1 and U_2 are discontinuous, their sum is not. I conclude that the proposition applies.

THEOREM 1: *Given* (1)–(5), *a mixed strategy equilibrium exists in the case of production in advance* (*where firm i's profit is given by* (6)).

I turn next to production to order. Here we encounter a difficulty in the application of the proposition; namely, the sum of profits need no longer be continuous nor even upper semicontinuous. Although discontinuities still involve a shift in demand between firms, they now also entail a shift in *production*. Thus an increase in production by a less efficient firm at the expense of a more efficient one may induce a fall in total profit.

To deal with this difficulty, I modify the strategy spaces and payoff functions somewhat to restore upper semicontinuity. This will enable us to conclude that an equilibrium for the modified payoff functions exists. I then argue that the strategies for this equilibrium remain in equilibrium for the original payoff functions.

I first strengthen the assumptions about cost functions. In addition to (1), we require

$$\text{(10)} \qquad c_i(x) \text{ is strictly convex.}$$

For $i=1,2$ firm i's supply function is

$$s_i(p) = \arg \max_{x \in [0, K_i]} [px - c_i(x)].$$

From (1) and (10), $s_i(p)$ is well-defined, nondecreasing, and continuous. Notice that by assuming that costs are convex, we can conclude that $s_i = s_i(p)$ maximizes

$\min\{d_i, s_i\}p - c_i(\min\{d_i, s_i\})$, for any d_i. Accordingly, in the modified model, let firm i's strategy space be $\hat{A}_i = [0, \bar{p}]$. For any price p, let $V_i(p) = px_i - c_i(x_i)$ and $W_i(p) = py_i - c_i(y_i)$, where $x_i = \min\{s_i(p), F(p)\}$ and $y_i = \min\{s_i(p), \max\{F(p) - s_j(p), 0\}\}$. Then,

$$(11)\quad W_1(p) \le U_1(p, s_1(p), p, s_2(p)) \le V_1(p);$$

$$(12)\quad W_2(p) \le U_2(p, s_1(p), p, s_2(p)) \le V_2(p).$$

The second inequality in (11) is strict if and only if the first inequality in (12) is strict. Moreover, the first inequality in (11) is strict if and only if the second inequality in (12) is strict. Thus we can choose $Z_i(p) \in [U_i(p, s_1(p), p, s_2(p)), V_i(p)]$ such that $Z_1(p)$ ($Z_2(p)$) is in the interior of the interval if the second inequality in (11) ((12)) is strict, and

$$(13)\quad Z_1(p) + Z_2(p) \ge \max\{W_1(p) + V_2(p), W_2(p) + V_1(p)\}.$$

If $\{(p_1^n, p_2^n)\}$ is a sequence converging to (p, p) then

$$(14)\quad \limsup \sum U_i(p_1^n, s_1(p_1^n), p_2^n, s_2(p_2^n)) \le \max\Big\{W_1(p) + V_2(p), W_2(p) + V_1(p), \sum U_i(p, s_1(p), p, s_2(p))\Big\}.$$

Define

$$(15)\quad \hat{U}_i(p_1, p_2) = \begin{cases} U_i(p_1, s_1(p_1), p_2, s_2(p_2)), & \text{if } p_1 \ne p_2 \\ Z_i(p), & \text{if } p_1 = p_2 = p \end{cases}$$

Combining (13)–(15), it can be deduced that $\Sigma \hat{U}_i$ is upper semicontinuous. If $p = 0$, $\hat{U}_i$ is continuous at p, and if $p > 0$, $\hat{U}_i$ is lower semicontinuous from the left. Hence, applying the above proposition (and simply ignoring the extra dimension of the strategy space), I conclude that the modified game has a mixed strategy equilibrium $(\hat{\mu}_1, \hat{\mu}_2)$.

I claim that $(\hat{\mu}_1, \hat{\mu}_2)$ also represents a mixed strategy equilibrium of the original game: if firm i plays p_i in the modified game, then it plays $(p_i, s_i(p_i))$ in the original game. Choose $\hat{p}_1$ in the support of $\hat{\mu}_1$. Then

$$(16)\quad \int \hat{U}_1(\hat{p}_1, p_2)\, d\hat{\mu}_2 \ge \int \hat{U}_1(p_1, p_2)\, d\hat{\mu}_2,$$

for all $p_1 \in [0, \bar{p}]$. If $\hat{\mu}_2$ places positive probability on $\hat{p}_1$ and $U_1(\hat{p}_1, s_1(\hat{p}_1), \hat{p}_1, s_2(\hat{p}_1)) \ne \hat{U}_1(\hat{p}_1, \hat{p}_2)$, then by (15) and from definition of $Z_1(\hat{p}_1)$ there exists $\tilde{p}_1$ slightly less than $\hat{p}_1$ such that $\int \hat{U}_1(\tilde{p}_1, p_2)\, d\hat{\mu}_2 > \int \hat{U}_1(\hat{p}_1, p_2)\, d\hat{\mu}_2$, a contradiction of (16). Hence, because $U_1(p_1, s_1(p_i), p_2, s_2(p_2))$ and $\hat{U}_1(p_1, p_2)$ differ only where $p_1 = p_2$, the left-hand side of (16) equals $\int U_1(\hat{p}_1, s_1(\hat{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2$. But, from (15), $\hat{U}_1(p_1, p_2)$ is greater than or equal to $U_1(p_1, s_1(p_1), p_2, s_2(p_2))$ everywhere, and so (16) implies

$$(17)\quad \int U_1(\hat{p}_1, s_1(\hat{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2 \ge \int U_1(p_1, s_1(p_1), p_2, s_2(p_2))\, d\hat{\mu}_2 \quad \text{for all } p_1.$$

Now suppose that for some p_1^0 and $s_1^0 \in [0, K_1]$,

$$\int U_1(\hat{p}_1, s_1(\hat{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2 < \int U_1(p_1^0, s_1^0, p_2, s_2(p_2))\, d\hat{\mu}_2.$$

Then, in view of (17), $U_1(p_1^0, s_1^0, p_1^0, s_2(p_1^0)) > U_1(p_1^0, s_1(p_1^0), p_1^0, s_2(p_1^0))$ and $\hat{\mu}_2(p_1^0) > 0$. Thus there exists $\tilde{p}_1$ slightly less than p_1^0 such that

$$\int U_1(\tilde{p}_1, s_1(\tilde{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2 > \int U_1(\hat{p}_1, s_1(\hat{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2,$$

a contradiction of (17). I conclude that

$$\int U_1(\hat{p}_1, s_1(\hat{p}_1), p_2, s_2(p_2))\, d\hat{\mu}_2$$
$$\geq \int U_1(p_1, s_1, p_2, s_2(p_2))\, d\hat{\mu}_2$$

for all p_1 and s_1.

I have demonstrated

THEOREM 2: *Given* (1)–(5) *and* (10), *a mixed strategy equilibrium exists in the case of production to order* (*where firm i's profit is given by* (8)).

REFERENCES

Allen, B. and Hellwig, M., "Bertrand-Edgeworth Oligopoly in Large Markets," mimeo., University of Bonn, 1983.

Beckmann, M., "Edgeworth-Bertrand Duopoly Revisited," in R. Henn, ed. *Operations Research-Verfahren, Vol. III*, Meisenheim: Sonderdruck, Verlag, Anton Hein, 1965, 55–68.

Bertrand, J., "Review of Cournot's 'Rechercher sur la theorie mathematique de la richesse'," *Journal des Savants*, 1883, 499–508.

Dasgupta, P. and Maskin, E., "Existence of Equilibrium in Discontinuous Economic Games, 1 and 2," *Review of Economic Studies*, forthcoming 1986.

Dixon, H., "Existence of Mixed Strategy Equilibria in a Price-Setting Oligopoly with Convex Costs," *Economic Letters*, 1984, *16*, 205–12.

______ **and Maskin, E.**, "The Existence of Equilibrium with Price-Setting Firms," mimeo., Harvard University, 1986.

Edgeworth, F., *Papers Relating to Political Economy*, London: Macmillan, 1925.

Fan, K., "Fixed Point and Minimax Theorems in Locally Convex Topological Linear Spaces," *Proceedings of the National Academy of Sciences*, 1952, *38*, 121–26.

Gertner, R., "Simultaneous Move Price-Quantity Games and Non-Market Clearing Equilibrium," mimeo., MIT, 1985.

Glicksberg, I., "A Further Generalization of the Kakutani Fixed Point Theorem with Application to Nash Equilibrium Points," *Proceedings of the American Mathematical Society*, 1952 *38*, 170–74.

Levitan, R. and Shubik, M., "Price Duopoly and Capacity Constraints," *International Economic Review*, February 1972, *13*, 111–22.

Simon, L., "Games with Discontinuous Payoffs, 1: Theory," mimeo., University of California-Berkeley, 1984.

Price-Setting Firms and the Oligopolistic Foundations of Perfect Competition

By BETH ALLEN AND MARTIN HELLWIG*

Two major objections can be made to the theory of perfect competition and its relevance to real markets. First, perfectly competitive theory assumes that individual economic agents have absolutely no market power. This idealization cannot be precisely true and, at best, can provide only an accurate and useful approximation. Second, the theory fails to explain how prices are formed. These observations lead us to study the theory of monopolistically competitive markets in which firms choose prices rather than quantities, and then to ask whether such large markets approximate perfect competition.

Price-setting oligopolies can provide economists with a price formation story; prices arise from the profit-maximizing decisions of individual firms. Thus, we are able to obtain a theory in which prices are picked by those economic actors who do so in reality. Although this is a matter of taste, we believe that the use of prices, instead of quantities, as the strategic variables of firms can yield a better simple descriptive model of oligopoly that captures the essential features of many markets. This motivation then leads to the question of which price-setting model should be used for the study of the oligopolistic foundations of perfect competition.

In the original model formulated by Joseph Bertrand (1883), firms have constant marginal and average costs so that any one firm can supply the whole market. This assumption leads to the conclusion that the presence of two (or more) price-setting firms suffices to yield perfectly competitive outcomes. Because of this result, the industrial organization and economics of information literature have utilized the Bertrand paradigm extensively. Thus, the standard models of "competitive" insurance, credit, and labor markets under asymmetric information all rely on the Bertrand approach.[1]

We have misgivings about this research strategy. In our view, the proposition that "two is competitive" does *not* in general provide a valid description of how markets work. We think the functioning of a market certainly depends on whether it contains 2 or 2000 firms.

Specifically, we believe that "perfect competition" requires the absence of market power. In the Bertrand model, firms have a lot of market power. Each firm can supply the entire market, thereby cutting out all other firms. In very simple models, this market power is completely neutralized by the particular strategic interactions among firms. However, the recent work of Marie-Odile Yanelle (1985) shows that in more complicated settings, the Bertrand model may lead to some very noncompetitive outcomes.[2]

*University of Pennsylvania, Philadelphia, PA 19104, and University of Bonn, West Germany, respectively. Our research on this topic was supported by the Deutsche Forschungsgemeinschaft through Sonderforschungsbereiche 21 and 303, by the National Science Foundation through research grant IST83-14096, and by a NATO Research Fellowship in Science.

[1]See, for example, Michael Spence (1973), Michael Rothschild and Joseph Stiglitz (1976), Charles Wilson (1977), and Stiglitz and Andrew Weiss (1981).

[2]Yanelle examines a Bertrand model of price-setting intermediaries who compete on an input market as well as an output market. There is a pure strategy equilibrium at Walrasian prices if and only if the Walrasian output price maximizes revenue. If the Walrasian output price does not maximize revenue, then there are two alternatives: (*i*) If firms choose only an output price and an input price as their strategic variables, then there is a pure strategy equilibrium in which the revenue maximizing (monopoly) output price is charged by all firms. (*ii*) If in addition to prices, firms can choose an upper bound on the aggregate input quantities that they will accept, then an equilibrium in pure strategies fails to exist.

The alternative that we propose is based on Francis Edgeworth's (1925) modification of the original Bertrand model. In this specification, firms have capacity constraints or strictly increasing marginal costs, which prevent any one firm from serving the whole market. We model the absence of market power by examining the limiting case where each firm is small relative to the market—that is, where capacity constraints or increasing marginal costs become effective long before the firm serves a significant portion of the market.

One drawback of the Bertrand-Edgeworth specification is that the noncooperative game among price-setting firms does not, in general, have an equilibrium in pure strategies. To ensure consistency of the theoretical model, firms must be allowed to randomize over a whole range of prices among which they are indifferent in equilibrium. However, our results show that this randomization is in some sense unimportant if the market contains many small firms.

More precisely, we find that if each firm is small relative to the market, then in a (mixed strategy) Bertrand-Edgeworth equilibrium, with high probability firms must charge prices close to a competitive price, and most transactions take place at prices near a competitive price. Thus the Bertrand-Edgeworth model provides us with an approximate account of competitive price formation which avoids the fiction of a Walrasian auctioneer.

I. Bertrand-Edgeworth Oligopoly

Consider the market for a single homogeneous good which is supplied by n firms.[3] Each firm $j=1,\dots,n$ has a quadratic cost $C_j^n(q)=q^2/2\gamma_j^n$, with $\gamma_j^n>0$, of producing an output q. If the output q is produced and sold at a price p, then firm j earns the profit

$$(1) \qquad \pi_j^n(p,q)=pq-q^2/2\gamma_j^n.$$

If firm j were a price taker in the market, then it would supply the profit-maximizing quantity $s_j^n(p)=\gamma_j^n p$ at the price p. Aggregate supply would then be given by $S(p)=Cp$ where $C=\Sigma_{j=1}^n\gamma_j^n$.

On the demand side, we assume a linear aggregate demand function $D(p)=A-Bp$, where $A>0$ and $B>0$. Hence there is a unique competitive price p_c at which aggregate demand and aggregate supply are equal so that $D(p_c)=S(p_c)$. There is also a unique monopoly price p_m which a monopolist or a cartel would charge in order to maximize the aggregate profit $pD(p)-\Sigma_{j=1}^n C_j^n(q_j)$, where $\Sigma_{j=1}^n q_j=D(p)$. One easily computes that $p_c=A/(B+C)$ and $p_m=A(B+C)/B(B+2C)$ so that, as one would expect, the competitive price is strictly less than the monopoly price.

However, there is no auctioneer or cartel to set prices. We assume that each firm j independently chooses a price p_j and a maximum quantity s_j that it is willing to sell at this price. These choices are taken by all firms simultaneously.

Consumers take the firms' choices as given and try to obtain the good as cheaply as possible. If the firms charging low prices have insufficient supplies, then consumers are rationed at these prices. This rationing takes the form of a queuing system so that at any firm, consumers in the front of the line are fully satisfied whereas consumers in the back of the queue get nothing and must go to another firm. (The set of all consumers is taken to be so large that we may neglect the one person in the queue who is only partially served.) Altogether this allocation of consumers to firms is taken to satisfy the following: (*i*) Lower-priced firms must sell out first before any sales are made by higher-priced firms. (*ii*) If two or more firms charge the same price, then they split their potential market randomly. (*iii*) Each firm serves a representative sample of customers and faces the same per capita demand as a function of price. If firm j's potential clientele is a fraction δ_j of the total population, then firm j faces the demand $\delta_j D(p_j)$ and makes the sale

$$(2) \qquad q_j(p_j,s_j,\delta_j)=\min(s_j,\delta_j D(p_j)).$$

[3] In contrast to the increasing marginal cost case analyzed here, our earlier papers (forthcoming, 1986; 1985) considered a model with fixed capacities. Even so, many of the details that we can only sketch in this paper can be filled in by referring to this earlier work.

Given p_j and s_j, firm j can serve up to a fraction $s_j/D(p_j)$ of the population. If δ_j exceeds $s_j/D(p_j)$, then the demand facing firm j exceeds its supply, and a fraction $\delta_j - s_j/D(p_j)$ of the consumer population is sent away to firms with higher prices. By the same reasoning, firm j's own clientele δ_j depends on the fraction $\Sigma_{p_i<p_j} s_i/D(p_i)$ of the population that could maximally be served by firms with prices below p_j. If this fraction exceeds one, then every consumer can be served at lower prices, and firm j has no customers. If, on the other hand, $\Sigma_{p_i<p_j} s_i/D(p_i) < 1$, then a fraction $1 - \Sigma_{p_i<p_j} s_i/D(p_i)$ of the consumer population is left over by the lower-priced firms. This clientele is split randomly among those firms k which charge the same price $p_k = p_j$. Denoting firm j's share as $\tilde{\lambda}_j$, we may write

$$(3)\quad \delta_j^n(p_1, s_1, \dots, p_n, s_n) = \max\left[0, \tilde{\lambda}_j\left(1 - \sum_{p_i<p_j} s_i/D(p_i)\right)\right]$$

for firm j's clientele in a market with n firms choosing $(p_1, s_1, \dots, p_n, s_n)$. Whereas $\tilde{\lambda}_j = 1$ if firm j is the only firm charging the price p_j, we assume that $\tilde{\lambda}_j$ is truly random whenever two or more firms charge the same price p_j. For the different firms k that charge this price, the random variables $\tilde{\lambda}_k$ must be interrelated in such a way that aggregate sales at p_j are equal to aggregate supplies at p_j or to residual aggregate demand at p_j, whichever is smaller.[4] Thus, we must have

$$(4)\quad \sum_{p_k=p_j} q_k(p_k, s_k, \delta_k^n(p_1, s_1, \dots, p_n, s_n)) = \min\left\{\sum_{p_k=p_j} s_k, D(p_j) \times \max\left[0, 1 - \sum_{p_i<p_j} s_i/D(p_i)\right]\right\}.$$

[4]This requirement implies that the random variables $\tilde{\lambda}_k$ depend on the choices $p_1, s_1, \dots, p_n, s_n$. Thus in the duopoly case, $\tilde{\lambda}_j$ might take the values 1 and $\max[0, 1 - s_k/D(p_k)]$ with probability one-half each.

We assume that production costs depend on actual sales rather than the chosen maximum supplies.[5] Then firm j's profits are given by

$$(5)\quad u_j^n(p_1, s_1, \dots, p_n, s_n) = \pi_j^n\big(p_j, q_j\big(p_j, s_j, \delta_j^n(p_1, s_1, \dots, p_n, s_n)\big)\big).$$

The profit functions u_j^n may be regarded as the payoff functions in a noncooperative game among firms. A noncooperative (Nash) equilibrium of this game is termed a Bertrand-Edgeworth equilibrium.

II. Properties of Bertrand-Edgeworth Oligopoly

Since the number of firms is finite, each firm in our model has market power. A firm can exploit those clients who cannot go elsewhere because the other firms' supplies are limited. It can also damage its competitors by undercutting them and taking away their customers.

For an illustration of the firm's monopoly power over its customers, consider firm j's position when all firms $k \neq j$ charge the competitive price p_c and offer the competitive supply $s_k^n(p_c) = \gamma_k^n p_c$. Since $D(p_c) = S(p_c) = Cp_c$, any firm $k \neq j$ serves at most a fraction $s_k^n(p_c)/D(p_c) = \gamma_k^n/C$ of the population. If firm j charges a price $p_j > p_c$, its clientele δ_j is equal to the remainder $1 - \Sigma_{k\neq j}\gamma_k^n/C = \gamma_j^n/C$, regardless of p_j. With this clientele, firm j earns the profit

$$(6)\quad \pi_j^n(p_j, D(p_j)\gamma_j^n/C) = \gamma_j^n\left[p_j D(p_j) - D(p_j)^2/2C\right]/C,$$

which is just the fraction γ_j^n/C of the aggregate profit that a monopolist or a cartel would earn at the price p_j. To maximize its profit in this situation, firm j should charge the monopoly price p_m, thereby exploiting its captive clientele in the same way that a monopolist would exploit the entire market.

[5]This assumption corresponds to the case of production to order. The alternative assumption that costs depend on maximum sales yields similar results.

To see the role of undercutting, consider a situation in which all firms charge the same price $p > p_c$. Then $D(p) < S(p)$, and at least one firm j must be selling less than its profit-maximizing supply $s_j^n(p)$. Trivally, we may also suppose that firm j sells less than $D(p)$. If firm j moves to a price p_j' slightly below p, then it undercuts the other firms and gets the whole market for a sale of the smaller of $D(p_j')$ and $s_j^n(p_j')$. Because the discrete jump in sales outweighs the marginal decrease in price, this move must raise firm j's profit.

More generally, suppose that two or more firms anticipate a tie (with positive probability) at the price p. Then unless they expect the tie to involve the profit-maximizing sales $s_k^n(p)$ for each of them, at least one firm j can increase its profits by moving to the price p_j' just below p and to the maximum supply $s_j^n(p_j')$.

These considerations show that price rather than supply is the important strategic variable in Bertrand-Edgeworth oligopoly. Indeed, we claim that in any Bertrand-Edgeworth equilibrium, each firm j may simply be taken to choose the profit-maximizing supply $s_j = s_j^n(p_j)$ that corresponds to the price p_j. In equilibrium, s_j is chosen solely for its direct impact on profits through the limitation of sales. Any effect of s_j on firm j's clientele δ_j is neglected because such an effect occurs only if s_j affects firm j's share $\tilde{\lambda}_j$ in the case of a tie. By the preceding argument, such a tie cannot occur (with positive probability) in equilibrium unless it involves the sale $s_j^n(p_j)$ and hence the supply $s_j = s_j^n(p_j)$.[6]

The prices charged in a Bertrand-Edgeworth equilibrium must be between the competitive price p_c and the monopoly price p_m.[7] At p_c, each firm j is always assured of selling its entire supply $s_j^n(p_c) = \gamma_j^n p_c$. Any firm k that might charge lower prices serves no more than a fraction $s_k^n(p_c)/D(p_c) \leq \gamma_k^n/C$ of the market so that firm j has at least the clientele $1-\Sigma_{k\neq j}\gamma_k^n/C = \gamma_j^n/C$ and the demand $D(p_c)\gamma_j^n/C = s_j^n(p_c)$. At p_c, firm j is thus sure to earn the profit $\pi_j^n(p_c, \gamma_j^n p_c)$, which is more than it can earn at a lower price. Therefore, prices below p_c are never charged.

Next we consider the firm j whose price p_j is the highest price in the market. Suppose that p_j exceeds the monopoly price p_m. We have seen that p_m would maximize firm j's profits from a clientele of size γ_j^n/C. Firm j's actual clientele δ_j cannot exceed $\gamma_j^n/C = 1-\Sigma_{k\neq j}\gamma_j^n/C$ because at prices above p_c, each firm $k \neq j$ can serve a clientele no smaller than γ_k^n/C. Since a smaller clientele involves smaller sales and hence a lower marginal cost of production, the profit-maximizing price for a clientele $\delta_j \leq \gamma_j^n/C$ cannot exceed p_m. A move from $p_j > p_m$ to p_m must therefore raise firm j's profits both because it brings the firm closer to profit maximization for the given clientele, and because it might increase firm j's clientele by undercutting other firms. Hence the highest price charged in equilibrium cannot exceed p_m.

However, as Edgeworth observed, there is no Bertrand-Edgeworth equilibrium in pure strategies—that is, for every constellation $(p_1, s_1, \ldots, p_n, s_n)$ of firms' decisions, there is a firm j that can improve itself by moving from (p_j, s_j) to some other choice (p_j', s_j'). Our previous arguments show that there is no equilibrium in which all firms charge the same price p. For $p > p_c$, some firm j would gain by undercutting, whereas for $p = p_c$, supplies would have to be $s_k^n(p_c)$ for $k = 1, \ldots, n$,[8] in which case any firm j would gain by moving from p_c to the monopoly price p_m. However, there is also no equilibrium in which two firms j and k charge

[6]The randomness of $\tilde{\lambda}_j$ in (3) rules out the possibility that $s_j > s_j^n(p_j) = \delta_j D(p_j)$.

[7]This assertion is invalid in more general models in which there is more than one competitive price. See our 1985 paper.

[8]This step of the argument and with it the claim of nonexistence of equilibrium in pure strategies would not be valid if the shares $\tilde{\lambda}_j$ in (3) were nonrandom (see fn. 6). Thus for $\lambda_j = \gamma_j^n/C$, the competitive price $p_j = p_c$ and supply $s_j = Cp_c$ yield a Bertrand-Edgeworth equilibrium.

different prices $p_k > p_j$. In such an equilibrium, the higher-priced firm k must have positive profits because it prefers p_k to the competitive price p_c. Hence firm k's sales must be positive, which means that firm j's supply constraint must be strictly binding. But then firm j must have a captive market in which case it can increase its profits by slightly raising its price without losing any customers.

Therefore we must allow for mixed strategies in which firms independently randomize over different prices. Eric Maskin (1986) shows that our market always possesses a Bertrand-Edgeworth equilibrium in mixed strategies. By our previous argument, this equilibrium involves prices between the competitive price p_c and the monopoly price p_m. In this interval, each firm is indifferent among the different prices that it chooses because each of them yields the same expected payoff when played against the equilibrium mixed strategies of other firms. The firm's uncertainty about the prices charged by other firms induces a perfect balance between the desire to raise the price in order to better exploit one's customers and the desire to lower the price in order to get more customers by undercutting the other firms.

III. Bertrand-Edgeworth and Competitive Equilibria

In this section, we argue that Bertrand-Edgeworth equilibria converge to perfectly competitive equilibria if the market becomes large and each firm becomes small relative to the market. For each $n = 2,3,\ldots,$ we consider a market with n firms having cost parameters $\gamma_1^n,\ldots,\gamma_n^n$ such that (i) for some constant $\Gamma > 0$, $0 < \gamma_j^n \le \Gamma/n$ for all n and every $j = 1,\ldots,n$, and (ii) for all n, $\Sigma_{j=1}^n \gamma_j^n = C$. The aggregate competitive supply function $S(p) = Cp$ is thus independent of n. The aggregate demand function $D(p) = A - Bp$ is also taken to be the same for all n. Therefore the competitive price p_c and the monopoly price p_m do not depend on n. We claim that if n is sufficiently large then, with probability close to one, most consumers make purchases at prices close to p_c and only a few are served at significantly higher prices. Similarly, with high probability, the fraction of firms that fail to sell their profit-maximizing supplies at nearly competitive prices becomes small. The Bertrand-Edgeworth equilibrium prices, aggregate transactions, and aggregate profits converge in distribution to the competitive price p_c, the competitive transaction $S(p_c) = D(p_c)$, and the competitive profit $p_c^2 C/2$.

The intuition for this result is partly based on the law of large numbers. The profitability of a given price p for firm j depends on the fraction of the population which is served at prices below p. Given the random variables $\tilde{p}_1^n,\ldots,\tilde{p}_n^n$ and $\tilde{s}_1^n = \gamma_1^n \tilde{p}_1^n,\ldots,\tilde{s}_n^n = \gamma_n^n \tilde{p}_n^n$ representing the prices and supplies chosen in a mixed strategy equilibrium with n firms, the fraction of the market which can be served by firms $i \neq j$ at prices less than p is

$$\tilde{\eta}_j^n(p) = \sum_{\substack{i \neq j \\ \tilde{p}_i^n < p}} \gamma_i^n \tilde{p}_i^n / D(\tilde{p}_i^n). \tag{7}$$

For firm j to make a sale at p, $\tilde{\eta}_j^n(p)$ must be less than one. Hence firm j charges the price p only if it expects $\tilde{\eta}_j^n(p)$ to be less than one sufficiently often. Observe that the prices charged by the different firms are bounded and mutually independent. Because the supply coefficients γ_i^n in (7) go to zero as n becomes large, the weak law of large numbers implies that for large n, $\tilde{\eta}_j^n(p)$ is unlikely to deviate significantly from its expectation $g_j^n(p) = E\tilde{\eta}_j^n(p)$. Thus, for sufficiently large n, firm j only charges the price p if the expectation $g_j^n(p)$ is not significantly greater than one, that is, if on average the potential supply by firms $i \neq j$ at prices below p is not much greater than demand. If $g_j^n(p)$ were greater than $1 + \varepsilon$ for some $\varepsilon > 0$ and very large n, then the probability that firm j can make a sale at p (i.e., the probability that $\tilde{\eta}_j^n(p)$ is less than one and hence less than $E\tilde{\eta}_j^n(p) - \varepsilon$) must be close to zero. In this case, firm j's expected profits at p must be less than what it obtains at the competitive price p_c.

We apply these considerations to the highest price $\bar{p}^n$ charged by any firm in a Bertrand-Edgeworth equilibrium with n firms and to the firm $j(n)$ that charges this price. At the price $\bar{p}^n$, firm $j(n)$ is always undercut by all other firms. Therefore $g^n_{j(n)}(\bar{p}^n)$ is simply the expected fraction of consumers that can be served by firms other than $j(n)$ at whatever prices they pick. The preceding argument implies that for any $\varepsilon > 0$, $g^n_{j(n)}(\bar{p}^n) = E[\Sigma_{i \neq j(n)}\gamma^n_i \tilde{p}^n_i / D(\tilde{p}^n_i)]$ cannot exceed $1+\varepsilon$ for arbitrarily large n. Since firm $j(n)$'s own supply $\gamma^n_{j(n)}\bar{p}^n_{j(n)}$ diminishes, the expectation $E[\Sigma^n_{i=1}\gamma^n_i \tilde{p}^n_i / D(\tilde{p}^n_i)]$ of the fraction of the population that can be served by all firms (including $j(n)$) converges to one as n becomes large. Because prices below the competitive price are never charged, the random variable $\Sigma\gamma^n_i \tilde{p}^n_i / D(\tilde{p}^n_i)$ is never less than $\Sigma\gamma^n_i p_c / D(p_c) = 1$, and we conclude that for large n, with probability close to one, most firms must charge prices close to the competitive price p_c.

We can interpret this result in terms of our previous discussion of undercutting vs. monopoly exploitation of captive clienteles. As n becomes large, each firm j's market share becomes small (regardless of price). If the other firms charge exactly the competitive price p_c, it still remains true that firm j should charge the monopoly price p_m. However, if the other firms charge prices above p_c and if firm j is small, then it should not move to a higher price because the other firms can seize its clientele. A small firm cannot afford to charge a high price unless all of the other firms charge prices very near p_c. Therefore as n becomes large, more and more weight is put on undercutting considerations whereas the monopoly exploitation of captive clienteles becomes less and less important. Thus in the Bertrand-Edgeworth approach, the absence of market power in the interaction of many small price-setting firms provides a foundation for perfectly competitive equilibrium.

REFERENCES

Allen, Beth and Hellwig, Martin, "Bertrand-Edgeworth Oligopoly in Large Markets," *Review of Economic Studies*, forthcoming 1986.

______ **and** ______, "The Approximation of Competitive Equilibria by Bertrand-Edgeworth Equilibria in Large Markets," Discussion Paper No. A-1, Sonderforschungsbereich 303, University of Bonn, January 1985.

Bertrand, J., "Review of 'Théorie Mathématique de la Richesse Sociale' and 'Recherches sur les Principes Mathématiques de la Théorie des Richesses,'" *Journal des Savants*, 1883, 499–508.

Edgeworth, F. Y., "The Pure Theory of Monopoly," in *Edgeworth*, *Papers Relating to Political Economy*, Vol. I, New York: Burt Franklin, 1925, ch. E.

Maskin, Eric, "The Existence of Equilibrium with Price-Setting Firms," *American Economic Review Proceedings*, May 1986, *76*, 382–86.

Rothschild, Michael and Stiglitz, Joseph, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics*, November 1976, *90*, 629–49.

Spence, Michael, "Job Market Signaling," *Quarterly Journal of Economics*, August 1973, *87*, 355–74.

Stiglitz, Joseph E. and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Wilson, Charles, "A Model of Insurance Markets with Incomplete Information," *Journal of Economic Theory*, December 1977, *16*, 167–207.

Yanelle, Marie-Odile, "Bertrand Competition Among Intermediaries," Discussion Paper A-31, Sonderforschungsbereich 303, University of Bonn, December 1985.

Vertical Product Differentiation: Some Basic Themes

By JOHN SUTTON*

If you ask most people why some firm has a large market share, you are likely to be told that it has "a great product," or words to that effect. That kind of explanation would strike an economist as being less than satisfying, however. There are two difficulties: Rolex watches, like IBM computers, may be an enviable product, but they aren't noted for their share of the watch market. Having a "better" product presumably means that the demand schedule faced by the firm is shifted further outwards, than would otherwise be the case. But the firm, so advantaged, might find it optimal to take its profit (largely) in the form of a higher price rather than in the form of an increased volume of sales. There is also a second difficulty: if this product is so successful, why don't more of its competitors develop similar products?

To elaborate on this second difficulty, we might argue as follows: if a large number of consumers are located in some region of "preference space," so that a sole producer in that region would enjoy a large market share and relatively high profits, then we would expect other firms to place themselves close by, and so erode the incumbent's advantaged position. Certainly, such a story is plausible within those traditional "location" models of the Hotelling type, which form the basis of much of the literature on "horizontal" product differentiation (in Figure 1, imagine $f(\alpha)$ to represent the density of consumers along Hotelling's line; the popularity of location A implies a greater density of suppliers in that area, rather than an advantaged position for some lucky incumbent A as against his rival B).

And yet, despite these difficulties there does seem to be something in the above naive explanation. To draw it out, however, we need to go beyond these location-type stories, in order to define what is meant by a better product. In so doing, we might gain some new perspectives on the well-trodden issue of why some industries are more concentrated than others.

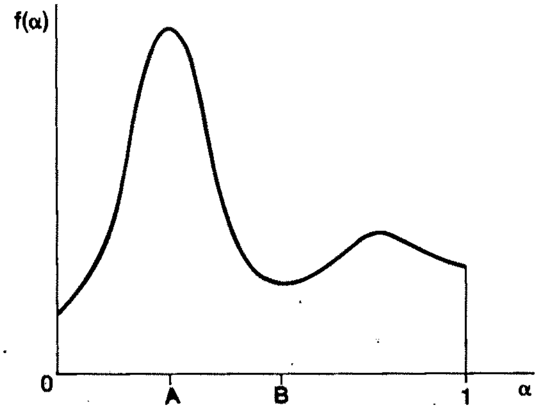


FIGURE 1

I. Vertical Product Differentiation

In the Hotelling-type model illustrated in Figure 1, if products A and B are offered at the same price, then each will have a positive market share.

This is not the case in models in which products are taken to differ in quality ("vertical" product differentiation). This latter kind of model has the defining property, that if two distinct products are offered at the same price, then *all* consumers prefer the same one (the higher-quality product). Now, of course, products will typically differ in respect of many attributes, some of which may be thought of as horizontal (the range of software offered with a computer, say) and some as vertical (its operating speed, say). It is helpful, however, to begin by looking at the "pure" case in which products differ in a single vertical attribute.

*London School of Economics, Aldwych, London WC2 2AE. The financial support of ICERD and the ESRC is gratefully acknowledged.

To provide a point of reference, consider again the simple location model of Figure 1, in which products differ in respect of a single horizontal attribute. Each consumer buys exactly one unit for which he pays the quoted price plus a transport cost, dependent on his distance from the producer. Each consumer buys from the "lowest-cost" source, and his tastes are fully described by his location on the interval.

Now ignore any sunk costs associated with entry for the moment, and imagine that each available product is produced at the same level of marginal cost c, and that this is constant for all levels of output (to emphasize the latter point, I label c as the unit variable cost). Suppose we make available, along this interval, any finite number of distinct products. Suppose each product is offered by a different firm, and suppose firms compete à la Bertrand, that is, we seek a Nash equilibrium in prices. Suppose such an equilibrium exists.[1] Then it is obvious that, at equilibrium, each firm has a positive market share, and a price exceeding unit variable cost; this follows from the fact that a firm can always capture consumers in its immediate neighborhood by offering a price just above unit variable cost;[2] and so has positive profits (ignoring any sunk costs incurred in entering the market). Now this property is of particular interest, for the following reason: suppose we introduce a "preceding" stage to this game, in which firms choose their respective products and incur some fixed cost in so doing. Equilibrium can then be characterized as a perfect equilibrium in this two-stage game. Now suppose that we increase the size of the market (by successive replications of the population of consumers); will the equilibrium configuration of market shares become more "fragmented," in the sense that the market shares of all firms become arbitrarily small?

[1] This will require restrictions on consumer preferences (the transport cost function), (see C. d'Aspremont et al., 1979; Damien Nevin, 1985).

[2] So long as one or more of its rivals are not selling at prices below unit variable cost; but this is excluded, by profit maximization.

The property of horizontal differentiation models noted above—that we can enter an unbounded number of firms, each earning revenue exceeding variable cost—is clearly a necessary (though by no means sufficient) condition for convergence to a fragmented structure. (To induce entry the margin earned at the second stage must suffice to cover the sunk costs incurred at the preceding entry stage.)

In models of (pure) vertical differentiation, this property may, or may not, hold good. Whether or not it does, depends on the nature of technology and tastes. Specifically, what matters is the relationship between consumers' willingness to pay for quality improvements, and the increase in *unit variable cost* associated with such improvements.

To fix ideas, consider the following example: let all consumers have identical tastes, viz, a consumer of income t derives utility $u\cdot(t-p)$ from consuming one unit of a product of quality level u, at price p; while if he buys none of the alternative brands of this good, his utility is represented by $u_0 \cdot t$. It is easily seen that, if a number of goods of various quality levels, $u_n > u_{n-1} > \dots > u_1 > u_0$ are offered at prices $p_n > p_{n-1} > \dots > p_1 > 0$, respectively, then consumers partition themselves by income, in such a way that brands of successively higher quality are purchased by consumers in successively higher income bands. (This reflects the fact that the utility function just defined, has the property that a consumer's willingness to pay for quality improvements is an increasing function of income.)

Suppose consumers' incomes are described by some density function $f(t)$ which takes a positive value on some interval $0 < a \leq t \leq b$. Denote as $c(u)$ the unit variable cost incurred in producing a product of quality u. Let $c(u)$ be defined on some feasible range $[u_0, \bar{u}]$, and suppose it is smooth. Consider the (hypothetical) case in which all qualities in $[u_0, \bar{u}]$ are made available at prices $p = c(u)$. Now suppose that, at these prices, richer consumers in our income interval $[a, b]$ prefer to purchase higher-quality products, than do poorer consumers. Then we can identify with each income level t a different preferred quality u: thus there is a quite precise corre-

spondence between this vertical differentiation model, and the horizontal case (where, in the location story, each consumer most prefers the product closest to him). In particular, the property noted earlier (that an unbounded number of firms can be fitted in with positive market shares and prices exceeding unit variable cost) holds good; and for exactly the same reasons as we noted above.

What is required for all this to hold good, is that $c(u)$ should increase steeply enough[3] with u on $[u_0, \bar{u}]$. If it does not, then it may be the case that all consumers prefer the same good (the highest quality offered). This is immediately obvious if $c(u)$ does not rise too steeply. Suppose then, that this is the case (i.e., at $p = c(u)$), all consumers rank the goods in the same order. Then it is no longer true that an infinite number of goods can be "fitted in"; instead, the following finiteness property holds: there exists an upper bound, independent of the qualities on offer, to the number of firms which can coexist with positive market shares, and a price exceeding unit variable cost, at a Nash equilibrium in prices. (See Avner Shaked's and my 1983, 1984 papers.)

This property is a very strong one; it implies a rather drastic failure of the tendency for market shares to become fragmented. The mechanism through which this effect operates is as follows: price competition between some group of high-quality products drives their prices down to a level at which even the poorest consumer prefers to buy one of these goods, at its equilibrium price, rather than buy any of the excluded goods, at a price sufficient to cover unit variable cost. If a new product of quality higher than those available is entered, the resulting repercussions on prices are such as to drive out some hitherto viable low quality rival(s) (J. Jaskold Gabzewicz and J.-F. Thisse, 1980). For an analysis of product choice in this setting, see Shaked's and my paper (1982).

What can be said then, of the relation between quality and market share across the various firms which survive? Here, no general relationship holds. The answer depends upon the distribution of consumers' willingness to pay. For a very broad class of such distributions, high-quality products tend to have a larger market share. To get a counterexample, we need to assume that consumer incomes (tastes) are highly skewed.[4] Indeed, it is intuitively clear that, if the distribution of income (or willingness to pay) has a long tail, then the top quality firm will find it optimal to take a small market share of high-income consumers, however low its unit variable cost.

The finiteness property just described, is a property which relates only to the latter (price competition) stage of the underlying game, and is independent of any considerations relating to optimal product choice. This in fact is a major simplifying feature in analyzing the pure vertical differentiation case. Once we go beyond this framework, and allow products to differ in both a vertical and a horizontal attribute, the property no longer holds. To see this simply, consider an (arbitrarily large) set of distinct products, all with the same vertical attribute. Then, as in the pure horizontal case, all will earn revenue exceeding variable cost at equilibrium.

It is natural to ask, however, whether such a configuration can be supported as an equilibrium? This brings us to a second, distinct, mechanism which can preempt convergence to a fragmented structure.

II. A Second Mechanism

Let us consider a model in which products differ according to two attributes, a horizontal attribute h and a vertical attribute u. A consumer's utility will now be represented by his income y (willingness to pay for quality improvement), and his most preferred value of h (called α). Moreover, suppose there is

[3] Of course, if $c(u)$ is too high, no product will be viable.

[4] For instance, in the case of the utility function cited above, and where $c(u)$ is constant for all u, then market shares are positively related to quality if the distribution of income is uniform, normal, or lognormal. It is difficult, however, to say what (utility function) is realistic here (Jaskold Gabszewicz et al. 1981).

some marginal rate of substitution between h and u, which is finite and positive. Suppose a firm incurs fixed cost $F(u)$ at stage I, and unit variable cost $c(u)$ at stage II, in making product (u, h).

In this setting, we can again pose the question: as the size of the market increases, can we converge to a fragmented market structure? The answer turns out to depend again on the nature of technology and tastes. A condition sufficient to exclude such convergence, can be stated loosely as follows (for a precise statement, the reader is referred to Shaked's and my 1985 paper):

(i) Unit variable cost should rise only slowly as quality increases;

(ii) The proportionate rate of increase in fixed cost associated with a given increase in quality (measured in terms of consumers' willingness to pay) should be bounded above, for all u.

The first of these conditions mirrors my earlier condition. (The second is required, because as the size of the economy increases in this setting, there will be a tendency for the spectrum of qualities on offer to rise in step; and so the relevant segment of $F(u)$ shifts upwards.)

Now the mechanism which prevents convergence to a fragmented structure, as the economy is replicated here, works as follows: convergence implies that for any ε we can find an economy size such that, once the economy exceeds this size, all firms must have market share less than ε at equilibrium. Denote the top quality on offer as u. Now suppose we can find an increment Δ, by which a firm can augment its quality u, thus jumping to $(u+\Delta)$. The fact that u and h are substitutes, permits it to capture a larger market segment, which includes consumers who prefer the horizontal attribute of some rival products. The question is whether capturing this larger market share will increase its profit. To ensure this, we need that variable costs should not rise too much (assumption (i)); we also need that the proportionate rise in its fixed costs should not be too great —but, in fact, it turns out that the boundedness condition (ii) is enough here. Under assumption (i) and (ii), then, this tendency for firms to jump to higher levels of u in order to escape "crowding" by competitors at lower levels, will suffice to exclude a fragmented structure.

An interesting feature of this second mechanism is that it seems to be fairly robust to alternative specifications of the underlying model (whether we assume simultaneous entry, or sequential entry; whether we use Bertrand or Cournot competition; whether firms are single product or multiproduct). Given the common complaint that "with oligopoly, anything can happen," this point is worth emphasizing. It should be remarked that while the characterization of equilibria indicated above is relatively straightforward, the existence of equilibria in these models, as in models of product differentiation generally, tends to be problematic. For a brief discussion of this, see Shaked's and my paper (forthcoming) and Tilman Borgers (1986).

III. Some Applications

Some of the attractions of these vertical product differentiation models, is that they provide a unified framework within which to explore issues that are typically addressed under the headings of *R&D*, or advertising. What is common to these areas, analytically, is the notion that enhanced expenditure on fixed cost may increase demand for the product in question, while there is no a priori reason to expect any particular rate of increase in variable costs.

An obvious area of application concerns the analysis of intra-industry trade in differentiated products. The case we have emphasized here (that in which the fragmentation result fails) is discussed in Shaked's and my paper (1984). In this setting, if we join two similar economies via international trade, then the result in the short run (where product specifications are fixed) tends to lead to a fall in prices of high-quality products, and the consequent exit of some low-quality producers. In the long run, however, the combined industry remains relatively concentrated, as the size of the market increases, and the returns to each firm from marginal improvements in quality are correspondingly greater—and so there is a tendency for the spectrum of qualities on offer to shift up-

wards. Thus the impact of trade, in these models, involves a further dimension to that observed in the (more widely studied) horizontal differentiation models.

Recent applications of these models involve the meshing of this literature with the literature on patent races, as in the work of J. Beath et al. (1985), while Y. Katsoulacos (1985) has examined the differing employment impact of product innovation, within the horizontal and vertical differentiation frameworks. As to extensions of the underlying analysis, some models which combine vertical and horizontal product differentiation have been studied by Shaked and myself (1985), and by Norman Ireland (1985).

Little has been done as yet on the analysis of competition between multiproduct firms offering differentiated products.[5] On this issue, the model of P. Champsaur and J. C. Rochet (1985) breaks new ground. They consider the "Chamberlinian" case of vertical differentiation—which is analogous to a horizontal differentiation model—and explore the process of competition between two firms offering (nonoverlapping) bands of products.

IV. Conclusions

The notion that having a better product is the key to acquiring a large market share, and to enhanced profitability, is more popular in business schools than in departments of economics (see for example, S. R. Schoeffler et al., 1974). Such a notion seems rather dubious if we think in terms of the standard horizontal product differentiation models. It is difficult to capture the concept of a better product within such a framework, however; the popular product at location A in Figure 1 above fails to fit the story. Once we admit vertical product differentiation, however, such a story makes more sense. In following through this idea, what appears to matter is the extent to which the burden of product improvement falls primarily on fixed costs, or on variable costs. Where the technology is such that product improvement can be achieved, through the expenditure of fixed costs, with a sufficiently small degree of increase in unit variable costs, then we are likely to find that industrial structure is relatively concentrated. I have set out two distinct mechanisms through which this effect may operate. What these two mechanisms have in common, is the notion that, loosely stated, a firm which can provide a product better in some regard, than those of its rivals, with a limited increase in its unit variable costs, can thereby capture a significant share of the market.

[5]Though see Shaked and myself (forthcoming). On the multiproduct monopolist, see Michael Mussa and Sherwin Rosen (1978), and Jaskold Gabszewicz et al. (1986).

REFERENCES

Beath, J., Katsoulacos, Y. and Ulph, D., "Sequential Product Innovation and Market Structure," Discussion Paper No. 85-171, Bristol University, 1985.

Borgers, Tilman, "Existence of Equilibrium with Sequential Entry in Product Differentiation Models," ICERD Working Paper, LSE, 1986.

Champsaur, P. and Rochet, J.-C., "Product Differentiation and Duopoly," unpublished working paper, 1985.

d'Aspremont, C., Jaskold Gabszewicz, J. and Thisse, J.-F., "On Hotelling's 'Stability in Competition'," *Econometrica*, September 1979, *47*, 1145–50.

Ireland, Norman, "Combining Horizontal and Vertical Product Differentiation," working paper, Warwick University, 1985.

Jaskold Gabszewicz, J., and Thisse, J.-F., "Entry (and Exit) in a Differentiated Industry," *Journal of Economic Theory*, April 1980, *22*, 327–38.

Jaskold Gabszewicz, J. et al., "Price Competition among Differentiated Products: A detailed study of a Nash Equilibrium," ICERD Discussion Paper, LSE, 1981.

______, "Segmenting the Market: The Monopolist's Optimal Product Mix," *Journal of Economic Theory*, forthcoming 1986.

Katsoulacos, Y., "Product Innovation & Employment," *European Economic Review*, February 1985, *26*, 83–108.

Mussa, M. and Rosen, S., "Monopoly and Product Quality," *Journal of Economic*

Work for Welfare: How Much Good Will It Do?

By FRANK S. LEVY AND RICHARD C. MICHEL*

It is an axiom of policy analysis that the good is in the particulars. The axiom is well-taken. When people discuss a new policy proposal in general terms, details are vague and the proposal can seem to serve a number of conflicting goals. It is only when we get down to details that the inherent conflicts and disappointments appear.

The axiom might usefully be applied to the increasingly frequent call for a tighter relationship between welfare and work. During the 1950's and 1960's, the idea was largely dormant. But over the last 15 years, it has gained wider acceptance for a number of different reasons. One is the feeling that requiring work of welfare recipients makes welfare more politically acceptable. A second is the sense that, contrary to social goals, welfare disconnected from work leads to a growing underclass within the poverty population (Charles Murray, 1984; Lawrence Mead, 1985).

More recently, a closer connection between work incentives and welfare has been implicitly advanced as the answer to a third problem: a looming budgetary crisis which has led to challenges to the cost of the American welfare state. This third rationale is the subject of this paper.

I. The Coming Crisis for the Welfare State

The emerging crisis for the welfare state in America begins with the changing nature of both poverty and the American income distribution.

Consider first the income distribution. From 1947 through the present, the bottom quintile of the income distribution has received a small and fairly steady share of all income—between 4.7 percent and 5.6 percent. But while the quintile shares of income have been relatively constant, the sources of this income have changed substantially. In 1959, the Census reported that about 70 percent of all income in the lowest quintile came from earnings. By 1969, the proportion had declined to 60 percent. And by 1983, earnings comprised only 42 percent of all income in the lowest quintile.[1] Government cash benefits have increased correspondingly with combined welfare and Social Security benefits approaching 45 percent of total income for this group.

The lowest quintile of the distribution and the poverty population are, of course, separate concepts: the first is relative while the second is absolute. But recent statistics on poverty show a comparable situation: in 1983, earnings comprised only 46 percent of the income of the poor and this occurred after a 2-year recession which had drawn many near (and presumably working) poor into the poverty population.

Government benefits thus now play a crucial role in shoring up the bottom of the income distribution. But it is precisely these benefits that are now called into question. The recently passed Gramm-Rudman legislation formalizes what most people long ago admitted: that the supply-side tax cuts of 1981 could never produce sufficient growth to become self-financing. They left a struc-

†*Discussants*: Gary Burtless, The Brookings Institution; David Ellwood, Harvard University.

*Professor, Public Affairs, University of Maryland, College Park, MD 20742; and Director, Income Security and Pension Policy Center, The Urban Institute, Washington, D.C. 20037, respectively.

[1]If we improved upon the Census definition of income to delete taxes and add the value of in-kind benefits such as Food Stamps, Medicaid, and Medicare, the proportion of this redefined income that came from earnings would undoubtedly fall still lower.

tural federal budget deficit that can only be closed by expenditure reductions or tax increases, and the reductions will apparently get first priority. For the moment, major income support programs such as Social Security and Aid to Families with Dependent Children are off limits in the sense that they are exempt from Gramm-Rudman's automatic reductions. But these automatic reductions are only a fail-safe strategy if Congress and the administration cannot agree on an alternate plan. In the initial phases of formulating such a plan, income support programs would surely be on the table, as they have been since 1981.

It seems then that the income distribution has evolved to the point where the goal of a balanced budget is on a collision course with the goals of reducing poverty and income inequality. In this situation, it is natural enough to look for escape hatches and implementing work requirements in welfare programs appears to provide one exit. If welfare recipients could be put into jobs—if earnings could substitute for benefits in their incomes —the collision described above might be avoided.

How might this substitution be accomplished? If the object is to save public funds, a large-scale public jobs program such as that proposed by the Carter Administration is obviously not the answer. What is implied is a gradual reduction in the value of welfare compared to the value of work: a tightening of eligibility restrictions and a continued erosion of maximum benefit levels by inflation which reduces real outlays. If—as the story goes—the expansion of the benefits increases dependency, then the contraction of benefits should decrease dependency. Earnings per family will increase and neither equality nor poverty will suffer. And ultimately, government outlays would be reduced.

The argument is appealing because in a pure accounting sense, it is clear that government benefits substituted for declining earnings. But it is essential to understand the causes of this substitution in order to determine whether a stronger work-welfare policy will succeed in saving substantial public monies. Only if the rising importance of transfers *caused* the replacement of earnings will this be true.

It is important to first understand that the shift in income sources among low-income families was accompanied by (and perhaps in part caused by) the growing relative numbers of poor female-headed families in the 1960's and 1970's. Between 1959 and 1984, the number of female-headed families in poverty grew from 1.9 million, making up 23 percent of all poor families to 3.5 million, making up 46 percent of all poor families. A corresponding dramatic growth in the number of families receiving welfare benefits in the late 1960's and early 1970's combined with the shift in the composition of the poor to make female-headed families the principal targets of government-initiated work incentive policies.

The question then remains as to what policies would reverse the processes by which female-headed families became poor in such large numbers and by which poor families came to rely less on earnings. The answer to this depends on what mechanisms led to these outcomes.

In practice, two such mechanisms are implicit in the literature. One is a labor supply effect in which transfers permit families to reduce work effort and earnings. The other is a demographic effect in which transfers induce the creation of female-headed families —families with traditionally weak earnings potential. One can make a theoretical case for either effect. But again the good is in the particulars and so it is necessary to assess how strong each effect is. In this paper, we review the evidence on each one: labor supply in Section III, family effects in Section IV, and in Section V, draw some brief conclusions.

II. Transfers and Labor Supply

In the short run— that is, the horizon over which we hope to balance the budget—it is far easier to imagine reversing labor supply effects than demographic effects and so it is with labor supply that we begin.

The case for labor supply effects begins with the observation of Murray and others that during the 1970's, *GNP* per capita grew rapidly—more rapidly than in the 1950's—but the poverty rate remained constant. Therefore, Murray argues, welfare programs

must have been luring people out of the labor force.[2] One might counter that this was really a demographic argument—that increasing benefits created more dependent households. But an overview of the postwar period suggests something more was going on. Let us define "dependent" families as families headed by someone over 65, or a woman under 65. This is an admittedly simple definition but one which corresponds to historically accepted definitions of groups which, when they are poor, deserve society's support.

During the postwar period, the proportion of such dependent families has grown in both the poverty population and the bottom quintile of the income distribution. But in both cases, the proportion of families with no member in the labor force has grown much faster (Table 1).

TABLE 1

A. Characteristics of the Lowest Quintile of the Family Income Distribution	1949	1979
Proportion of families headed by a person over 65 or a woman under 65	50%	59%
Proportion of all families with no earner	21%	40%
B. Selected Characteristics of the Poverty Population	**1959**	**1981**
Proportion of families headed by a person over 65 or a woman under 65	41%	57%
Proportion of all families with no earner	24%	39%

Sources: Herman Miller (1966), Bureau of the Census *Current Population Reports*, Series P-60, No. 147, and our tabulations from Decennial Census files (1960, 1970).

A closer inspection of the data shows that to the extent that labor has been cut back, it is not on the part of female family heads. Among those females heading families in poverty, the proportion who worked at some time during the year has fallen only slightly over time: 43 percent in 1959 and 1969, 37 percent in 1975, 39 percent in 1981 and 38 percent in 1984. The proportion who worked year-round behaved in a similar fashion declining from 16 percent in 1959 to 13 percent in 1969, and then dipping slightly in the 1970's only to return to 13 percent in 1984. The small magnitude of potential labor supply effects have been confirmed by a wealth of micro-level analysis as well. While the effect is almost always significant in a statistical sense, it rarely reaches proportions large enough to have more than nominal budgetary impacts (Levy, 1979, and Robert Moffitt, 1984.)

A further look at welfare benefits helps explain the stability in the proportion of workers. For many years, researchers argued that the work incentives of welfare were contained in its tax rates. In practice, however, this view overstated the case. Unlike a standard income tax, welfare has only a one-month accounting period. Over so short a period, there is little room to modulate labor supply, particulary for low-income workers who typically have no control over their hours of work for pay. Thus, an AFDC mother is generally faced with a decision to work full time for a period or not at all. And if she chooses to work full time, she is likely to be off the welfare roles altogether, if only for that month.

Given this kind of on-off behavior, one simple measure of welfare's work disincentives is the gap between welfare benefits and wages. Let us approximate the value of work by using the average wage paid in the retail trade industry.[3]

[2]The argument itself is fallacious on a number of counts. First, per capita income was not a good measure of economic growth in the 1970's because it was largely a result of demographic adjustments. Second, the link between poverty and welfare that Murray makes implies that leisure was so valuable to these families that they opted for the subpoverty benefits in AFDC over minimum wage jobs. But between 1967 and 1973, welfare roles (as measured by participation in the AFDC program) *grew* by 5.9 million persons while the number of persons below the poverty line *fell* by 4.8 million persons. For detailed arguments, see, among other, our paper (1985).

[3]This industry generally includes checkout clerks in a variety of retail stores such as groceries and fast-food outlets and has historically had the lowest hourly wage of major nonagricultural industries. In 1984, retail trade wages were about two-thirds of manufacturing wages and less than one-half of construction wages.

TABLE 2—COMPARISON OF COMBINED WELFARE BENEFITS WITH RETAIL TRADE WAGES

Year	Average Annual AFDC plus Food Stamp Benefit	Full-Time Annual Wages	Ratio of Welfare to Wages
1960	1,269	3,162	.401
1970	2,880	5,075	.567
1975	3,315	6,989	.474
1980	4,333	10,150	.427
1983	4,741	11,939	.397

Sources: We calculated welfare and food stamp benefits from program data. Food stamp benefits are those for a family with average AFDC income. Wages are from the *Economic Report of the President*, February 1985, Table B-38, page 276.

Looking at the value of the combined AFDC and food stamp benefits to an average AFDC family over time and comparing it to full-time work in the retail trade industry shows that while the relative generosity of benefits rose during the 1960's, it peaked in 1970, and by 1983 had returned to 1960 levels. In 1960, welfare benefits stood at 40 percent of wages. They rose to 57 percent by the early 1970's and declined to 40 percent in recent years (Table 2).

Throughout this period, the estimated participation rate in AFDC fluctuated. In the late 1960's, less than one-half of all eligible female-headed families participated in the AFDC program. This percent rose to 92 percent in the early 1970's, fell again in the mid-1970's, then reached a new peak of 97 percent in 1979. Recent calculations indicate that the rate has dropped steadily since 1979, reaching 78 percent in 1983 (Michel, 1980, and unpublished data from the Urban Institute's TRIM2 simulation model).

Furthermore, during this same period, the labor force participation rates among all women (white, black, poor, and nonpoor) continued to rise. Between 1970 and 1983, labor force participation rate of white women rose from 43 percent to 53 percent, duplicating the percentage changes for all women. For black women, the participation rate rose from 50 to 54 percent and for poor female family heads from 47 to 54 percent.

If in fact a rise in welfare benefits causes a large reduction in labor effort among women, there is no evidence in either aggregate or micro data. Whether welfare benefits were rising or falling, whether participation rates in welfare increased or decreased, and whether poverty rates were growing or shrinking, the labor force participation rates of the largest group categorically eligible for welfare continued to rise over the period from the late 1960's to the early 1980's.

To the extent that the growth in aggregate government transfers had any important labor supply effects, it was of course among the elderly. Throughout the postwar period (and in fact much earlier), the labor force participation among elderly men (those 65 and older) has fallen steadily from about 50 percent at the end of World War II to less than 20 percent today. A principal reason for this may have been the maturation of the Social Security program which allowed many men to leave the labor force in their 60's and still have an adequate income. Reflective of this is the fact that in 1979, only 25 percent of all men over 65 reported any earnings and only 8 percent reported year-round full-time work. Not surprisingly then, 42 percent of all families headed by an elderly family were in the lowest quintile of the income distribution.

It seems to follow from the numbers above that if we seek to limit or reverse the growth of the postwar transfer system by curbing labor force effects of transfer programs, the place to concentrate is not among female-headed families but among the elderly. No one has seriously proposed workfare for the elderly (nor are we), but it is important to understand that in terms of pure work incentive effects, that is where the dollars are.

III. Demographic Effects

If welfare payments do not have strong labor supply effects, what are their effects on family structure? The argument we advance is similar to the one advanced by black writers beginning with W. E. B. DuBois (1899) and Franklin Frazier (1939). It has most recently been articulately revived by William Julius Wilson (1978). If a man cannot find suitable work and has few prospects for finding work, then it is logical to ask what he can bring to a marriage and, in

particular, if he brings more income to a marriage than available from welfare. Thus if welfare benefits are higher than the incomes of a significant portion of men, it may provide an incentive to create more female-headed families.

Evidence from Decennial Census data and the March *Current Population Survey* files is suggestive (though by no means conclusive). In 1960, the average welfare benefit for an AFDC family was $1,269. About 31 percent of black males aged 20–24 had incomes below this level. By 1970, the average combined AFDC and food stamp benefit was $2,880. According to the *CPS*, fully 46 percent of young black men aged 20–24 had incomes lower than that figure. Combined welfare benefits also exceeded the incomes reported by 17 percent of black men aged 25–34 (compared with 15 percent in 1960).

Earlier, we saw how the gap between welfare and wages reached a minimum in the early 1970's and increased thereafter. But, while welfare benefits declined, the bottom part of the distribution of black men's incomes declined as well. In 1983, the average combined welfare benefit was $4,741. This figure was higher than the incomes reported in that year for 62 percent of black men aged 20–24 and 29 percent of black men aged 25–34. Both figures represent dramatic increases from earlier periods.

The potential relationship of those figures to the family formation prospects of black males and females can be demonstrated by linking them to the cohort sizes of males and females of marriageable age. In 1960, for example, there were 1.1 million black men and 1.3 million black women aged 25–34, a ratio of 1.2 women for every man. The cohort size of both sex groups nearly doubled by 1983 and this ratio remained roughly the same. But, if we look at the ratio of black women in this age group to black men whose income was above average welfare benefits, the ratio rises from 1.3 in 1960 to 1.6 in 1983. Similar increases in the ratio were experienced for younger black men and women. These facts confirm the findings of Wilson, and suggest a situation in which there was a decreasing number of black men who could provide income significant enough to maintain a family at above welfare level.

In presenting these findings, we recognize the counterargument: if women did not have the welfare option, more husbands might be "acceptable" and, given the responsibility, husbands might indeed work more. This is certainly a possibility, but the logical policy alternative here is to eliminate welfare for single women altogether. In spite of the fact that some conservatives have suggested this, it is not a policy likely to be adopted if only because current average welfare benefits are about one-half the poverty line.

IV. Conclusion

In Section I, we showed that government transfers had replaced earnings as the most significant source of income for the poor in the last 15 years. It seems that this replacement has occurred through two different mechanisms. First, the elderly have reduced their labor supply participation substantially. Second, the growth of transfers combined with market conditions or behavioral changes of unknown origin has induced the formation of female-headed families by pricing potential husbands out of the family market.

Having made these assertions, it is not clear how these problems can be resolved given current budgetary pressures and the apparent disinclination of the public to expand programs for the poor. Workfare for the elderly is almost certainly a nonstarter politically. And while one might want to put welfare mothers to work, work requirements alone are unlikely to solve what appears to be a major source of the welfare problem: the financial barriers to the formation of two-parent families. Workfare for AFDC mothers may be good policy for other reasons, but most certainly it will not provide significant near-term budgetary savings, since it is not labor-leisure decisions but demographic ones that are acting to maintain the size of the eligible population. The thing to do may be to find a way to put young men to work, but most policy options are too expensive to be considered seriously in this decade.

We began this paper with one axiom of policy analysis that the good is in the particulars. The particulars in the case of welfare policy have in recent years led to a stalemate. Changing the system invariably would in-

volve cutting it back to levels that appear to be socially unacceptable or expanding the system at costs which are currently unaffordable. And in recent years, humanitarian instincts have clashed with fiscal constraint to prevent us from moving in either direction.

REFERENCES

DuBois, W. E. B., *The Philadelphia Negro* (1899), rev. ed., New York: Schoken Books, 1967.

Frazier, Franklin E., *The Negro Family in the United States*, Chicago: University of Chicago Press, 1939.

Levy, Frank S., "The Labor Supply of Female Household Heads, or AFDC Work Incentives Don't Work Too Well," *Journal of Human Resources*, Winter 1979, *4*, 56–79.

______ **and Michel, Richard C.**, "Losing Perspective: The Recent Debate Over Welfare and Poverty," Working Paper 2081-02, Urban Institute, June 1985.

Mead, Lawrence W., *Beyond Entitlement: The Social Obligations of Citizenship*, New York: Free Press, 1985.

Michel, Richard C., "Participation Rates in the Aid to Families with Dependent Children Program: National Trends From 1967 to 1977," Working Paper No. 1387–02, Urban Institute, December 1980.

Miller, Herman P., *Income Distribution in the United States*, Washington: USGPO, 1966.

Moffitt, Robert, "Assessing the Effects of the 1981 Federal AFDC Legislation on the Work Effort of Welfare Recipients." Discussion Paper 742–84, University of Wisconsin Institute for Research on Poverty, January 1984.

Murray, Charles, *Losing Ground: American Social Policy, 1950–1980*, New York: Basic Books, 1984.

Wilson, William J., *The Declining Significance of Race: Blacks and Changing American Institutions*, Chicago: University of Chicago Press, 1978.

U.S. Bureau of the Census, *Current Population Reports*, Series P–60, 1973–83.

______, Decennial Census files, 1960, 1970.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, 1985.

Do Rising Tides Lift All Boats? The Impact of Secular and Cyclical Changes on Poverty

By SHELDON DANZIGER AND PETER GOTTSCHALK*

Discussions about the antipoverty effects of economic growth in the United States have largely been predicated on John Kennedy's metaphor that a rising tide lifts all boats. But the magnitude of these effects has been a subject of debate since the inception of the War on Poverty (see Lowell Gallaway, 1965, and Henry Aaron, 1967). This debate has public policy as well as academic implications—the greater the antipoverty effectiveness of growth, the less the need for special programs or income supplements during economic expansions.

Elsewhere, we have shown that increased real income need not be associated with a decline in poverty (see our 1984 and 1985 papers). In fact, poverty rates did not fall from 1982 to 1983, even though real median income increased. And in 1984 the official poverty rate was about the same as it was in 1967, while real median family income was 7.1 percent above its 1967 level.[1] If a rising tide was lifting all boats, the tide was late in many harbors.

In this paper we examine the relationship between macroeconomic conditions and poverty. Section I argues that several factors now limit the effectiveness of growth in reducing poverty. Section II differentiates the effects of secular economic growth from those of cyclical recoveries. The next section presents our interpretation of the data, followed by a brief conclusion. We show that growth had a large antipoverty effect through the early 1970's, but that the more recent experience has been different because growth rates have slowed and inequality has increased.

I. Factors Limiting the Antipoverty Effectiveness of Economic Growth

A. *Demographic Factors*

While an improvement in macroconditions can raise the income of poor households with an able-bodied head, it alone cannot eliminate poverty for households whose heads have weak attachment to the labor force. There are simply too many poor households that cannot benefit directly from improved labor market conditions.

In 1979, the last cyclical peak, almost two-thirds of all poor households were headed by a person who was elderly, a student, disabled, or a woman with a child under age 6. Given today's social norms, these heads of household can be classified as not expected to work. Indeed, almost all of them did not work during this year of relatively low unemployment. While these families will not gain directly from growth, they may benefit indirectly if a portion of the increased tax revenues resulting from growth are distributed through antipoverty programs.

The proportion of poor households not expected to benefit from economic expansion is not only large, but growing. In 1939, when poverty rates were much higher, less than one-third of poor household heads were classified as not expected to work by our defini-

*Director, Institute for Research on Poverty, University of Wisconsin, Madison, WI 53706, and Research Affiliate, Institute for Research on Poverty and Professor of Economics, Bowdoin College, Brunswick, ME 04011, respectively. Rebecca Blank, Elizabeth Evanson, and Lawrence Summers provided helpful comments on a previous draft; Steven Berry and George Slotsve provided computational assistance.

[1]These increases in real median family income and poverty rates between 1967 and 1984 are reported by the Census Bureau and use the Consumer Price Index (*CPI*). If one uses the *CPI-X* instead (which uses a rental equivalence approach to the cost of homeownership), the 1984 poverty rate was about 11 percent below the 1967 rate, while mean income was about 17 percent above its 1967 level. Use of the *CPI-X* does not alter the pattern of results shown in Table 1 below.

tion. From a purely demographic standpoint, it was easier to reduce poverty through growth of the economy in the 1940's and 1950's than it is today.

B. *Changes in the Shape of the Distribution*

The second factor limiting the antipoverty effectiveness of improved macroconditions is that the shape as well as the mean of the income distribution may change. Poverty status is determined by comparing a household's real income to its poverty line, which is fixed in real terms. This definition translates easily into standard statistical concepts—poverty is simply the cumulative density of income up to the poverty line. Therefore, poverty reflects changes in both position and shape of the distribution.

Most previous research and policy discussions have focused on the antipoverty impact of changes in mean income, implicitly assuming that only changes in the first moment of the income distribution affects poverty. Clearly, changes in the variance, skewness, kurtosis, and higher-level moments also influence poverty.

Development economists have shown that industrialization may increase the mean but also increase inequality, so there is no a priori reason to believe that the poor will benefit from growth. The empirical question is whether improved economic conditions in an advanced industrialized country are necessarily linked to stable or growing inequality. If inequality increases, then by definition a rising tide will not lift all boats in an equal manner.

C. *Nonlinear Relationship Between Growth and Poverty*

The third factor limiting the impact of growth on poverty is that as long as the poverty line falls to the left of the mode of the income distribution, fewer and fewer people will be taken out of poverty as the distribution shifts to the right. This results from the decreasing density of the distribution as one moves away from the mode. For example, suppose that all incomes increased by 2 percent a year for several years. All those households with incomes within 2 percent of the poverty line would exit poverty in each successive year, but the number leaving poverty would diminish each year. Thus, even if there were no changes in the demographic composition of the poor or in the shape of the income distribution, there would still be diminishing returns to economic growth.

II. Relationship Between Secular Growth, Cyclical Recoveries, and Poverty

We distinguish between the impact of secular and cyclical changes in macroconditions because the two reflect different underlying processes. While it is often difficult to make this distinction empirically, the conceptual difference is important. Secular growth reflects increases in potential output arising from increases in productivity and increases in factor supply. On the other hand, cyclical improvements operate primarily by increasing the quantity of labor employed. To argue that recovery from recession is an important antipoverty weapon is very different from arguing that policies which encourage growth through investment or technological change will reduce poverty substantially.

There is broad agreement that improved cyclical conditions are crucial in reducing poverty among households with an able-bodied member. Decreases in unemployment tend to increase mean income, decrease inequality, and reduce poverty (Rebecca Blank and Alan Blinder, 1986; Martin Dooley and Gottschalk, 1984).

The impacts of secular economic growth on poverty are less certain for two reasons. First, the yearly secular growth rate of the mean is generally small compared to cyclical increases, because long-term growth operates primarily through higher wage rates—over a one-year period it is much easier to increase the number of hours an unemployed person works than it is to increase the productivity of a fully employed person.

Second, growth may be accompanied by increases in inequality. Technological change and increases in physical and human capital will increase the average productivity of fully

employed persons. However, it does not follow that all wages will increase at the same rate. Technology or the structure of demands for final output may change in such a way that the earnings of low-skilled workers may even decline.

III. Empirical Patterns

A. *Secular Growth*

Using data from the *Census of Population Reports*, 1950 through 1980, we first review the relationship between poverty and the location and shape of the distribution. Since the effects of cyclical swings become less important over these 10-year periods, we interpret these changes in poverty as reflecting secular changes. Because 1969 and 1979 were both cyclical peaks, this interpretation is particularly appropriate for the last decade covered.

We analyze data for households headed by men aged 25 to 64. Relative to other households, they have the strongest attachment to the labor force and the smallest reliance on income transfers.[2] They are, therefore, the group most likely to benefit directly from economic growth. Our focus on prime-aged men also provides a rough correction for demographic change by excluding households headed by women, who have above-average poverty rates and represent an increasing proportion of all households. We do not deal with the possibility that more rapid economic growth could reduce this trend toward female headship.

Because the poverty line varies with family size, poverty will fall if family size declines, even if household income is constant, *ceteris paribus*. To control for the decline in family size that has occurred, we analyze the ratio of a household's income to its poverty line (the income-to-needs ratio).

TABLE 1—SECULAR GROWTH AND THE TREND IN POVERTY FOR HOUSEHOLDS HEADED BY MEN AGED 25–64, 1949–79

	1949	1959	1969	1979
Mean (Income/Needs)[a]	1.600	2.408	3.330	3.789
Variance ln(Income/Needs)[a]	.738	.771	.627	.730
Poverty Rate[b]	33.1	16.2	7.7	7.1
Change in Poverty Rate Due to:[c]				
Change in Mean	–	−13.8	−6.1	−1.3
Change in Shape	–	−3.1	−2.4	+0.7
Percentage Point Decline in Poverty Due to a 1 Percent Increase in the Mean, Holding Inequality Constant:[d]	−0.77	−0.28	−0.12	−0.09

Source: Computations from computer tapes of the *Census of Population Reports*, 1950, 1960, 1970, and 1980.

[a]Because the poverty lines are adjusted for changes in the *CPI*, income/needs ratios are fixed in real terms. We measured poverty in 1949 by adjusting the official lines back from 1959 using the *CPI* in the same way that they have been brought forward to the present.

[b]Shown in percent.

[c]The percentage-point difference between the poverty rates over any decade equals the sum of the fourth and fifth rows in the column for the latter year.

[d]Defined as the percentage point difference between the actual poverty rate in each year and a rate that results from increasing each household's income-to-needs ratio in the base year by 1 percent.

The first three rows of Table 1 show the mean and log variance of the income-to-needs ratio and the poverty rate in each of the Census years.[3] The log variance is an inequality measure which is sensitive to changes in the lower tail of the distribution. Note, however, that it measures only one aspect of change in the shape of the distribution, since

[2]The 1940 Census and the March 1968–March 1985 *Current Population Surveys* contain the information needed to classify household heads as expected and not expected to work. However, the 1950 Census does not provide information on students or the disabled, so our samples in Tables 1 and 2 differ. In 1979, those expected to work constituted about 70 percent of household heads, while male household heads between 25 and 64 years of age were 55 percent of all heads.

[3]Our measure of secular growth reflects changes in the heads' earnings, in the earnings of other family members, and in nonearned income, because poverty is officially measured on the basis of household income. Thus, the reduction in poverty resulting from the declines in family size and increases in income transfers that have occurred are attributed to secular growth, and inflate our estimates of the antipoverty effects of growth. The decline in poverty resulting from the increased labor force participation rate of wives is appropriately captured by our estimate of secular growth.

it does not reflect changes in third and higher-level moments.

The third and first rows of the table show that poverty declined when the mean increased and that the declines in poverty and the increases in the mean become successively smaller with each passing decade. At this superficial level, it seems that a rising tide was indeed lifting all boats. Such bivariate relationships do not, however, hold other factors constant.

We decompose the observed change in poverty over each decade into one component associated with shifts in the mean, and another with changes in the shape. The following thought experiment illustrates this decomposition. First, suppose that every household experienced the average increase in income in relation to needs. There would be no change in inequality and, as the distribution shifted to the right, poverty would drop. The difference between the initial-year poverty rate and this simulated rate gives the change in poverty when inequality is held constant.

Second, the effects of changes in inequality are obtained by comparing this simulated distribution to the actual distribution in the later year. By definition, the means of the two distributions are the same. However, if the actual distribution is less equal than the simulated distribution, changes in the shape will have increased poverty. By definition, the actual change in poverty over the decade is the sum of these two partial effects.

The fourth row of Table 1 shows how poverty rates would have changed if all households had experienced the average growth in the income-to-needs ratio. A rising mean was the primary cause of the reduction in poverty over the 30 years. However, the antipoverty effect of growth in the mean decreased in each successive decade (fourth row), primarily because of the falling rate of secular growth (first row).

In addition, the antipoverty effect of growth declined because of the nonlinear relationship between growth and poverty. The sixth row shows the percentage point decline in the poverty rate associated with a 1 percent increase in the mean, holding inequality constant. As poverty declined from 33.1 to 7.1 percent between 1949 and 1979, this measure of the antipoverty effect declined from -0.77 to -0.09 percentage points. Thus, a given percentage increase in the mean removed a much smaller number of households from poverty as the poverty rate declined.[4]

The fifth row shows the impact of changes in inequality, holding the mean constant. The changes in poverty due to changes in inequality were much smaller than those due to growth in the mean (compare the fourth and fifth rows). However, between 1969 and 1979, two years of comparable unemployment rates, the change in the shape of the distribution was poverty increasing, and offset roughly half of the poverty-decreasing effect of the rising mean.

The rise in inequality since 1969 has been well-documented but there is no consensus on its causes. Dooley and Gottschalk showed that increasing inequality reflects more than demographic or cyclical changes since inequality increased for males even after controlling for unemployment rates, education, experience, and growth in cohort size. Labor supply responses to transfers may be a partial explanation but they cannot explain the observed increases in inequality among prime-age males who are not well-covered by transfers. Likewise, changes in the occupational structure of the population are not sufficient to explain the changes in inequality found within occupational groups by Peter Henle and Paul Ryscavage (1980). Just as the slowdown in economic growth is not fully understood, neither is the rise in inequality. Both, however, occurred at about the same time and served to limit the secular decline in poverty during the 1970's.

[4]If a distribution is unimodal, a constant absolute increase in the mean will by definition yield a declining percentage point change in poverty. The sixth row shows that a constant percentage increase in the mean also has a declining impact. In fact, the elasticity—the percentage decline in poverty with respect to a constant percentage increase in the mean—also declines (data not shown).

Extrapolations of the future antipoverty effects of secular growth are quite difficult. If the increased inequality was caused by the same factors that in recent years have reduced the rate of productivity growth, and hence economic growth, then a resumption of productivity growth might reverse the trend toward inequality. However, productivity growth could increase without affecting the extent of inequality. It is this case, coupled with uncertainty about whether greater growth rates can be achieved, that makes us skeptical about the efficacy of relying primarily on economic growth to reduce poverty.

B. *Cyclical Changes*

Yearly microdata are available only since 1967, a period which covers three recessions. We use these data to assess the impact of cyclical conditions on poverty.

Table 2 shows the mean and log variance of the income-to-needs ratio, the poverty rate, and our decomposition of the change in poverty for the cyclical peaks and troughs between 1969 and 1982. To control for growth in the proportion of households not expected to benefit from cyclical improvements, we limit the analysis to households with heads who were under 65 years of age and were not disabled, students, or women with children under age 6. These "expected-to-work" heads composed about 70 percent of all household heads and about one-third of poor household heads for each of the years shown.

Column 1 shows that the mean exhibited a strong cyclical pattern around a small upward trend—it fell during each recession and surpassed the previous peak during each recovery. However, the 1975 and 1982 recessions were unusually deep and the intervening recovery was not strong.[5]

TABLE 2—CYCLICAL CHANGES AND THE TREND IN POVERTY FOR HOUSEHOLDS HEADED BY PERSONS EXPECTED TO WORK, 1969–82[a]

Year	Mean and ln Variance of Income Needs[b] (1)	Poverty Rate (2)	Change in Poverty Rate Due to Changes in:[c] Mean (3)	Shape (4)
1969: Peak	3.192 [.504]	6.9	–	–
1970: Trough	3.185 [.547]	7.3	+0.1	+0.3
1973: Peak	3.527 [.533]	6.3	−1.3	+0.3
1975: Trough	3.334 [.587]	7.8	+0.7	+0.8
1979: Peak	3.619 [.603]	6.8	−1.1	+0.1
1982: Trough	3.559 [.773]	10.4	+1.0	+2.6

Source: Computations from data tapes of March *CPS* 1970, 1971, 1974, 1975, 1980, and 1983.

[a]We define household heads as "expected to work" if they are younger than 65, not students, not disabled, and not women with children under age 6. Cols. 2–4 are shown in percent.

[b]Figures in brackets represent ln variance of income needs. Because the poverty lines are adjusted for changes in the *CPI*, income/needs ratios are fixed in real terms.

[c]The percentage-point difference between the poverty rates in any 2 years equals the sum of cols. 3 and 4 in the row for the latter year.

Inequality (shown in brackets in col. 1 for each year) exhibited small cyclical changes relative to a strong upward trend. Inequality increased not only between peaks and troughs, but also during the 1975 to 1979 recovery.

The last two columns show our decomposition of changes in poverty over the business cycle. The impact of cyclical changes in the mean (col. 3) is apparent—poverty varies countercyclically when inequality is held constant. Changes in the shape of the distribution (col. 4) increased poverty during all recoveries as well as recessions.[6]

[5]Table 2 does not show the data for 1984 because that year was not a peak. However, the most recent data follow the patterns discussed. The mean in 1984 of 3.592 was above the 1982 trough, but slightly below the 1979 peak; the log variance declined from the 1982 level to .758, still well above the 1979 level.

[6]While our Census and *CPS* samples differ, the results in Tables 1 and 2 for the 1969–79 period are similar: poverty declines slightly and the poverty-reducing effect of the rising mean is partially offset by the

IV. Conclusion

We have emphasized that there are limits to the antipoverty effects of improved economic conditions. First, because only about one-third of poor households have heads who are expected to work, most poor households will not benefit from improved labor market conditions. Therefore, we limited our empirical work to households with an able-bodied head to focus on those most likely to be affected by changes in macroconditions.

Second, we argued that the effects of a rising mean could be offset by increases in inequality. The empirical results show that prior to 1969 this was not the case, but that rising inequality during the 1970's and 1980's was a major factor contributing to increases in poverty.

The third limitation on the effect of macroconditions is that growth has a diminishing impact as poverty rates fall—since fewer people are near the poverty line, many fewer are pulled over the line by equal percentage increases in the mean. We found this effect to be quantitatively important.

Economic growth has been the primary source of poverty reduction in the past. However, in the absence of an unexpected increase in the rate of economic growth or an unforeseen decline in inequality, it seems unlikely that growth will substantially reduce poverty in the near future.

poverty-increasing effect of the changing shape of the distribution. For the entire 1969–82 period in the *CPS* sample, the poverty-increasing impact of growing inequality fully explains the observed increase in poverty.

REFERENCES

Aaron, Henry, "The Foundations of the War on Poverty Reexamined," *American Economic Review*, December 1967, *57*, 1229–40.

Blank, Rebecca and Blinder, Alan, "Macroeconomics, Income Distribution, and Poverty," in Sheldon H. Danziger and Daniel H. Weinberg, eds., *Fighting Poverty: What Works and What Doesn't*, Cambridge: Harvard University Press, 1986, 182–208.

Dooley, Martin and Gottschalk, Peter, "Earnings Inequality among Males in the United States: Trends and the Effect of Labor Force Growth," *Journal of Political Economy*, January 1984, *92*, 59–89.

Gallaway, Lowell, "The Foundations of the War on Poverty," *American Economic Review*, March 1965, *55*, 122–31.

Gottschalk, Peter and Danziger, Sheldon, "Macroeconomic Conditions, Income Transfers and the Trend in Poverty," in D. Lee Bawden, ed., *The Social Contract Revisited*, Washington: Urban Institute Press, 1984, 185–215.

______ and ______, "A Framework for Evaluating the Effects of Economic Growth and Transfers on Poverty," *American Economic Review*, March 1985, *75*, 153–61.

Henle, Peter, and Ryscavage, Paul, "The Distribution of Earned Income among Men and Women, 1958–77," *Monthly Labor Review*, April 1980, *103*, 3–10.

LECTURE ON ECONOMICS IN GOVERNMENT†

An Economic Accountant's Audit

By GEORGE JASZI*

When I was invited to speak at this meeting, I decided to audit my work as an economic accountant. What I shall say is drawn from my experience, of about 20 years as a staff member and about 23 as Director, of what is now the Bureau of Economic Analysis.

Let me first introduce this organization. The BEA constructs the U.S. economic accounts: it puts together a complex jigsaw puzzle—for which the pieces are the data collected mainly by the Census, BLS, Treasury, and OMB—that depicts the economic process. (BEA is much smaller than its sibling organizations, BLS and Census, because, except in the international field, it collects few source data.) The National Income and Product Accounts (*NIPAs*) provide a bird's-eye view of the economic process. They are supplemented by accounts showing tangible reproducible wealth. International transactions are seen in detail in the balance of payments and associated investment accounts. The input-output accounts show how industries interact to produce *GNP*. The regional information system, although far short of a set of regional accounts, is made up of small building blocks—the statistics for the more than 3,000 counties. The environmental accounts record the costs of abating pollution.

The BEA prepares a consolidated saving-investment account in which financial transactions necessarily cancel out. The Federal Reserve Board does the work on financial saving and investment and on financial assets. Other kinks in boundaries at times give rise to border conflicts. In the productivity ratio, BEA is the main source of the numerator, and BLS is the main source of the denominator; the intimacy of this Siamese-twin relationship can be grating. Another border conflict is between the Fed and BEA. The Fed's Index of Industrial Production and the comparable components of *GNP* often point in opposite directions.

The BEA also does forward-oriented work: a quarterly econometric model, balance of payments forecasts, regional projections, the system of cyclical indicators, and surveys of capital spending plans. Finally, BEA uses its economic accounts, the forward-oriented tools, and other information to analyze economic developments.

The users and uses of the work of BEA range widely. An important user is the public: business, labor, universities, state and local governments, research institutions, and informed individuals. An equally important user is the federal government: within Commerce, the Under Secretary for Economic Affairs; outside, the CEA, Treasury, OMB, and the Fed.

Over the 43 years at BEA, I feel that I made my principal contribution as an economic accountant. First, I resisted the will-o'-the-wisp of forging national output into a measure of economic welfare. I was a minority of one in a company that included such mental giants as Simon Kuznets and John Hicks, and at one point I had to defy a forceful Secretary of Commerce who had instructed the BEA to prepare a measure of welfare. Second, I helped construct the accounting system that describes the economic process, an enterprise that required perseverance, craftsmanship, a sense of the beauty of the system, and the ability to listen

†Sponsored jointly by the AEA/SGE.

*Chevy Chase, MD. I received in-depth assistance preparing this paper from Carol Carson and Helen Jaszi and valuable comments from Robert Parker, Hans Landsberg, Allan Young, Charles Waite, and Herbert Zassenhaus.

responsively.[1] I will elaborate impressionistically.

I struggled first with the several definitions of national output that dominated the field. Kuznets had defined national income as the "...net total of desirable events enjoyed by individuals..." (1933, p. 1). I perceived that, equipped with such a definition, no one could possibly measure it. It dawned on me that as a first step in bringing Kuznets' definition down to earth, "goods and services bought" must be substituted for "desirable events enjoyed." Gradually, I realized that the product attributable to a business unit can be defined only in terms based on business accounting concepts. Specifically, the value of production equals sales plus inventory change minus current-account purchases. Also, after consolidation of the units (and some additions for nonbusiness production and for product and income in kind), *GNP* in terms of product flows is the sum of sales to consumers, gross private capital formation, net exports, and sales to government. This definition is not a will-o'-the-wisp. It has stood the test of time. Furthermore, I saw national income as the sum of costs and profits that originate in the production of national output.

An accounting system that provides an overview of the economic process then fell into place. The system shows how *GNP* is distributed to consumers, business investors, foreign nations, and government. It shows how incomes that originate in the production of *GNP*, as modified by taxes and transfers, flow to these groups, and how they allocate these flows between consumption, and saving and investment. This overview makes it an unrivaled tool for macroeconomic analysis.

The overview has always been to me a thing of beauty, much like the view from a plane. But I must return to earth. Even though national output cannot be forged into a measure of welfare, it is of close relevance to it. To assess welfare, one must know, for instance, what part of *GNP* goes to national defense and what part to civilian programs, and whether an increase in *GNP* is "real" or due to inflation. Also, the measurement of national output must stay firmly anchored to business accounting, even though most of the working life of economic accountants is devoted to departing from it. For instance, they reject the methods used by most of business to value inventories and depreciation, and they devise radically different methods to serve their own purposes.

The meaning of the responsive listening I have mentioned is not obvious. Let me illustrate. The very idea of income and product accounting, which emerged in the early 1940's, grew out of the practical needs of economic mobilization for World War II and made its debut almost simultaneously in the statistical offices here and in Canada and England. A great deal of responsive listening was taking place, and it continued. By the mid-1950's, it was widely recognized that alternative systems, such as flow of funds and financial balance sheets as well as input-output, might be useful for policymaking, and there was great concern that these systems should be integrated with the *NIPA*s. By the 1960's, policy focused on economic stabilization, mainly through fiscal policy. More detailed and timely estimates of government transactions were required, as were more timely estimates of the *NIPA*s. Concurrently, the need for adjusted budget measures (the Model T of which was the full-employment budget) became evident. In the late 1960's and in the 1970's, high rates of inflation began to interfere with the functioning of the economy, and improved and new information on prices was required. Growth of productivity retarded sharply in the 1970's, and with it came special interest in investment, depreciation, and capital stocks. In the 1980's, concern with the underground economy burgeoned. The BEA listened responsively to these developments and produced the bulk of the information needed for policymaking associated with them. (See my papers with Carol Carson, 1981 and 1985.)

Next I shall take up the major issues I faced. The direction the work program should take was not one of them. When I became Director of the BEA, I believed that I should reorient the program. Subsequently I was

[1] See my 1971 article, Edward Denison (1971), and Arthur Okun (1971).

convinced that this was unnecessary—all I had to do was to do a little better what had been done before.

How the production and distribution of the economic accounts should be organized is an important issue. The source data used to estimate the entries in the accounts are whenever possible the by-products of administrative programs such as unemployment insurance, income tax, and the U.S. budget. General purpose statistics prepared by Census are the main source for all components of the product side of the *GNP* account, with the exception of federal purchases. Because these data differ in coverage and definition from the entries in the accounts, ingenious procedures that adjust the data must be devised. Reliance on the by-products of administrative operations secures inexpensively a large quantity of high-quality data.

Given the inherent shortcomings of the decentralized statistical system, strong coordination is required to ensure adequate coverage and definitional consistency in the economic accounts. After a history of generally unsuccessful coordination, the Statistical Policy Office, headed by the Chief Statistician, has been buried in OMB and its staff reduced to a handful. From the standpoint of constructing a solid set of accounts, it is most urgent to strengthen this group substantially (see Katherine Wallman et al., 1983).

The conversion of the data into estimates presented no difficulties because I always could rely on unusually competent estimators. But writing up BEA's work for the *Survey of Current Business* was difficult. I insisted on clear writing, both because most of the work of BEA reaches users through the *Survey*, and because of the close link between clear thinking and writing. In the words of Ben Jonson: "Neither can his mind be thought to be in tune, whose words do jarre, nor his reason in frame, whose sentence is preposterous."

The *Survey* has been a clearly written and informative publication—increasingly so since the mid-1970's. But, looking back, I feel that I should have tried harder to recruit a staff of manuscript editors to assist the editors-in-chief. The way in which the job is organized violates the principle of comparative advantage.

I want to comment on the long-overdue project of describing the concepts, definitions, and methodology underlying the *NIPA*s. An elegant introductory article introducing the concepts and definitions appeared in the March 1985 *Survey*, an admirable paper documenting the estimates of corporate profits became available in May, one documenting international transactions is due out in the spring, and others are in the pipeline. I had intended to publish all parts simultaneously, but deep in my heart I knew that the project would never see the light of day. To avoid this impasse, the new management chose serialization, even though it would not solve every problem (such as obsolescence, omissions, duplication, inconsistencies, etc.). This formidable list might well describe the obstacles that Charles Dickens faced in serializing his novels. He surmounted them except for *The Mystery of Edwin Drood*, which he never finished. My concern is that the BEA project may share its fate.

Given the imperfections of the source data, statistical error in the economic accounts is unavoidable. It is important to assess how accurate the estimates are and how accurate they should be. But I want to stress that accuracy is not the only standard by which the usefulness of the *NIPA*s should be judged. Even if the *NIPA* system consisted of only empty boxes, it would be useful to a wide range of economists—Keynesians, monetarists, supply siders, and rational expectationalists. The concepts, definitions, and classifications that make up the system have become a common language and have replaced the Babelian confusion that existed before the formulation of the system.

In this paper, accuracy is assessed by the standard method: if no revision is made in an estimate, it is considered to have no error; any difference between an estimate and a later revised one is considered to be the error. To a large extent, the method has merit: because later estimates are usually based on more source data, they are closer to the "true" number. The problem is the impossibility of determining how close any

estimate, including the final estimate, approaches the unknowable "true" number.

Several *NIPA* estimates are made for each quarter. The minus 15-day ("flash") estimates, prepared about two weeks before the end of the quarter and based on source data for only one or two months of the quarter, are released only in summary form: current- and constant-dollar *GNP*, the *GNP* implicit price deflator, and the *GNP* fixed-weighted price index. The 15-day estimates, which are based on two or three months' data, are shown in nearly 50 tables containing tremendous detail about the economy. The 45-day and 75-day estimates, which incorporate further data, are made available in the same detail. The estimates usually are revised further each July to incorporate annual data, and comprehensive revisions are made every 5 years largely to incorporate data from the economic and other censuses.

The minus 15-day and the 15-day estimates have been called the "eighth wonder of the world." However, those who object to the extent to which the estimates are based on assumptions take a negative view. It would be useful if a standard of accuracy could be specified, but unfortunately, it cannot. Desirable accuracy depends on the views one takes on several tradeoffs: the need for timely estimates; the relative importance of aggregates, overviews, and detail; and the relative importance of extending, as opposed to refining, the economic accounts.

As to timeliness: the minus 15-day and 15-day estimates have gaps in the two most volatile components of *GNP* (change in business inventories and net exports). Accordingly, accuracy could be improved significantly only by scrapping these estimates. The sacrifice of timeliness would be substantial: the 45-day and 75-day estimates, which are free from these statistical blemishes, are much too late to be used in short-run decision making.

As to aggregates vs. overview: accuracy in aggregates should not have highest priority. *GNP* should not be regarded as a stop-watch, but rather as a part of an overview of the economic process.

As to overview vs. detail: the main purpose of the *NIPA*s is to serve macro- rather than microeconomic analysis. For example, for a number of quarters in the early 1970's, estimates of total consumer expenditures for goods were fairly accurate, but the component estimates of food and consumer durables were off by billions: the benchmark estimates for food had been carried forward by the sales of food stores, which had added baby carriages, bicycles, and other durables to their lines. However, macroeconomic decisions were not impaired by this compositional error.

As to extensions vs. refinement: I fervently favor extensions. For instance, I think that even rough order-of-magnitude estimates of the illegal underground economy, which is not measured as part of the economic accounts, would be vastly preferable to most refinements.

My less than stringent view about accuracy stems from these four tradeoffs. However, the *NIPA*s have not measured up even to those views. At times they misled both macroeconomists and policymakers. I readily admit that it would have been preferable if these instances had not occurred. Of the nine instances enumerated by responsible critics, I have placed only three in my *NIPA* chamber of horrors. 1) During 1965, the early estimates failed to indicate the extraordinary strong expansion of the economy and the need for more restrictive economic policies. 2) In July 1971, corporate profits for 1969 and 1970 were revised down sharply; the initial estimates had understated their drop during the 1969–70 recession. 3) The inventory buildup during 1973 was vastly understated in the initial estimates, which thus provided a falsely optimistic view of economic developments. Had the buildup been registered, the 1974–75 recession might have been anticipated.

Using a different approach, I note five instances in 1968–83 in which large revisions in real *GNP* coincided with directional misses in critical quarters of the business cycle. As of now, I am placing none of them in my chamber of horrors. Three of them occurred in 1981–82; this period will be affected by the comprehensive revisions of the *NIPA*s that is being completed. Another instance is the second quarter of 1975; this is not really

a directional miss, because the 15-day estimate shows no directional change in that quarter. The fifth instance—the fourth quarter of 1969—puzzles me; I would have expected it to draw unfavorable comment, but it did not.

I should like to draw some practical conclusions. First, a sustained effort is needed to teach users that the *NIPA*s are eminently useful in macroeconomic analysis if they are not regarded as a precision instrument and that they may be lethal if they are. I remember with dismay being told by a distinguished Secretary of the Treasury that a complex set of fiscal, monetary, balance-of-payments, and wage-price policies, which determined the course of the U.S. economy for several years ahead, might not have been adopted had he known that *GNP* would be revised upward 1 percent. Although his view would have appalled me in any event, my shock was intensified because I knew that the revision stemmed from the incorporation of some delayed declarations of foreign dividends, which are a part of the net export component of *GNP*. Second, the minus 15-day *GNP* estimates are released only in the form of four summary aggregates. The major components of these estimates are not significantly less accurate than the components of the 15-day estimates, and I continue to advocate publication of all major components of the minus 15-day estimates. Withholding them is another instance of the over-precision syndrome. Third, nothing I have said should be taken to mean that I do not care about improved and new data. On the contrary, I have fought hard to obtain them. Over the past 8 years or so, restrictions on statistical budgets have gone about as far as they can go without seriously affecting the quantity, quality and timeliness of the *NIPA*s. Equally important, these restrictions have become an impediment to estimates and analysis.

I turn now to statistical discrepancies, a particularly obtrusive manifestation of the error in entries in the accounts: they are differences between two aggregates that would be equal if the estimates replicated underlying accounting relationships. The question is: should discrepancies be given free rein, be reduced by the exercise of judgment, or be eliminated? Each quarter BEA contends with two current-dollar discrepancies.

The first is that in the balance of payments accounts. In principle, the balance on current account should equal, with opposite sign, the balance on capital account. Nevertheless, a discrepancy appears because the methodologies used to estimate the two accounts are independent. In recent years, this discrepancy has shot up to astronomic heights. It is given free rein. I agree with the BEA experts that most of the discrepancy, and particularly quarterly fluctuations in it, stems from errors in the capital account. However, other experts have disagreed. Moreover, an IMF study of the global current-account discrepancy may suggest that part of the discrepancy is due to error in the U.S. current account. Finally, BEA has mounted a large-scale investigation of U.S. foreign trade in services, partly to find out whether or not certain exports that have grown rapidly in recent years are understated. Because of these uncertainties, giving the discrepancy free rein is the only course.

The second discrepancy is the *GNP* account, the two sides of which should in principle be equal. Nevertheless, a discrepancy appears because the methodologies used to estimate its income and product sides are largely independent. In this instance, the discrepancy is reduced by the exercise of judgment mainly because intensive study of these methodologies often provides solid clues for inferring which components are responsible.

The Cambridge (England) Group has proposed a mathematical method for eliminating discrepancies from the economic accounts. The discrepancy is eliminated "...with the aid of a generalized least squares algorithm for adjusting national accounts with subjective estimates of reliability of the various account items." In the light of this quotation, it is not clear to me that the mathematical method is less judgmental than the BEA method. Obviously, the Cambridge method would be preferable to giving discrepancies free rein if it were to provide estimates closer to true values, and in similar

circumstances would be preferable also to the judgmental method (see Terry Barker et al., 1984, p. 461).

I now turn to politicization. Contrary to occasional suspicion, there never has been an attempt to tamper with BEA's estimates. Not only would tampering have been risky, but a hands-off course was ensured because every administration needs the best possible figures. However, new administrations may also want new interpretations. Organizations doing economic analysis are at risk because politically displeasing analysis can be thwarted by restrictive budgets and personnel ceilings. It might appear that amputation of analysis could have eliminated all such risks. However, amputation was never a practical alternative for BEA. It would have seriously impaired BEA's estimates, which have benefited in a major way from internal use. Furthermore, those who oversee BEA would not have permitted amputation because of their reliance on BEA for analysis.

I always was aware of the politicization risk and long ago concluded that the safe course was to keep away from policy advice. This formula was my guide: BEA analyzes past economic events, tries to spot emerging problems, and will formulate alternative solutions without a recommendation. Making recommendations is left for policymakers and their advisers. This formula has worked, and I trust will continue to do so.

Even before I retired, the impression was abroad that I was a watchdog on the alert to prevent politicization at BEA. Quite to the contrary, the formula worked, and so did I —instead of having to bark or to bite. For this I am truly grateful.

You will have gathered that, on the whole, I regard my career as having been useful and rewarding. At times, however, I have been plagued by doubts. I will say what these doubts are and what affirmations I have marshalled. My doubts have to do with the relation of economic accounting to forecasting. I often have said that the accounts would be useless if they were relevant only to the past and present. Their sole justification is that they help forecast economic events. Policymakers and their advisers should be interested, for instance, not in whether there was or is inflation, but in whether there will be inflation. Many economists take a similar view. Forecasts—even the nonpartisan sort prepared by BEA—are subject to huge margins of error. Moreover, I believe that forecasting error is likely to increase. I know that increasingly sophisticated ways of specifying, estimating, and solving models are helping to reduce error. However, I believe that these improvements will be swamped by developments that will increase it. It is likely to become more and more difficult to foresee exogenous factors such as international economic developments, wars, and the risk of wars. In these circumstances, should increased reliance be placed on simulation? Simulations will not solve the problem, because it may become increasingly difficult to predict the set of endogenous relationships that characterize an economy.

Here is a rundown of my affirmations. Let us assume first that forecasting is indeed the sole justification for economic accounting. The error to which the forecasts are subject may not seem so disabling if we substitute the more relaxed phrase "learning from the past" for the tighter phrase "forecasting the future." In addition, some important forecasts are safe: It is just as certain that the distribution of income will not change overnight as that the sun will rise in the morning.

Consider next that forecasting may not be necessary to make economic accounting useful. In the natural sciences, no one questions the usefulness of the theory of evolution just because it does not tell us about future evolution. For economic history, a similar statement can be made.

Also, if we regard economic accounting as an art, nothing is wrong with taking pleasure in having painted a good picture of the economy. This is a highly subjective affirmation to which I am deeply attached.

About the validity of my final affirmation, I have no doubt. It relates to the importance of clear thinking. In the course of my professional life, I have had an opportunity to devote much of my time to the clarification of concepts, and I have concluded that it has profound practical—including moral—im-

plications. Confucius explains this in the following quotation in which he is the "Master," and Tse-Lu is an assistant, perhaps a policy adviser:

> Tse-Lu said: "The Prince of Wei is ready to hand over the reins of government to you. What is the first task that you will undertake, Master?" The Master said: "Unquestionably, the clarification of concepts." Tse-Lu said: "How impractical you are. Why care for the clarification of concepts?" The Master said: "How crude you are, Tse-Lu."...
>
> If concepts are not clear, words do not fit. If words do not fit, the day's work cannot be accomplished. If the day's work cannot be accomplished, morals and art do not flourish. If morals and art do not flourish, punishments are not just. If punishments are not just, the people do not know where to put hand or foot. [*Analects*]

In my clarification of concepts, I could not have done without my associates, and I am deeply grateful to them.

REFERENCES

Barker, Terry et al., "A Balanced System of National Accounts for the United Kingdom," *Review of Income and Wealth*, December 1984, *30*, 461–85.

Carson, Carol S. and Jaszi, George, "The National Income and Product Accounts of the United States: An Overview," *Survey of Current Business*, February 1981, *61*, 22–34.

______ **and** ______, "The Use of National Income and Product Accounts for Public Policy: Our Successes and Failures," paper presented at Joint Statistical Meetings, Las Vegas, NV, August 1985.

Denison, Edward F., "Welfare Measurement and the GNP," *Survey of Current Business*, January 1971, *51*, 13–16, 39.

Jaszi, George, "An Economic Accountant's Ledger," *The Economic Accounts of the United States, Prospect and Retrospect*, *Survey of Current Business*, July 1971, Part II, *51*, 225–27.

Kuznets, Simon, "National Income, Concepts, and Measurements," mimeo., submitted to *Encyclopedia of Social Sciences*, April 1933.

Okun, Arthur, "Social Welfare Has No Price Tag, *The Economic Accounts of the United States,...*, *Survey of Current Business*, July 1971, Part II, *51*, 129–33.

Wallman, Katherine K. et al., "Federal Statistical Coordination Today: An Epilogue as Prologue," *American Statistician*, August 1983, *37*, 177–202.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
NINETY-EIGHTH
ANNUAL
MEETING

NEW YORK, NEW YORK
DECEMBER 28–30, 1985

THE JOHN BATES CLARK AWARD

Citation on the Occasion of the Presentation of the Medal to

JERRY A. HAUSMAN

December 29, 1985

Jerry Hausman is an extraordinary applied econometrician. His innovative use of econometric methods and economic theory in public finance, labor economics, and energy has changed empirical work in these fields. He has developed techniques that are now in the toolbox of every applied economist: Tests for the exogenicity of right-hand side variables in a regression, exact welfare calculations from econometric demand functions, and the treatment of nonlinear budget constraints arising from income tax and income maintenance programs. His empirical findings are provocative and illuminating. He has challenged conventional wisdom on the incidence of taxes on labor supply and on savings and investment decisions. His finding of high discount rates in consumer appliance purchase decisions has forced a reevaluation of the economic effects of mandatory efficiency standards. His interweaving of economic theory and statistics to form powerful tools for analysis of economic problems is an example to us all.

Minutes of the Annual Meeting
New York, New York
December 29, 1985

The ninety-eighth annual meeting of the American Economic Association was called to order by President Charles Kindleberger at 5:55 p.m., December 29, 1985 in the New York Hilton Hotel. Before proceeding to items of business on the agenda, he commended Moses Abramovitz, the retiring editor of the *Journal of Economic Literature*, and Robert Clower, the retiring editor of the *American Economic Review*, for their outstanding work on the journals. There followed an enthusiastic round of applause for the two.

The minutes of the December 29, 1984 meeting were approved as published in the *American Economic Review, Papers and Proceedings* (May 1985, p. 417). The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Managing Editor of the *American Economic Review* (Orley Ashenfelter), Managing Editor of the *Journal of Economic Literature* (Abramovitz), and Director of *Job Openings for Economists* (Hinshaw) briefly reviewed their written reports which were available at the meeting. (See their reports published elsewhere in this issue.)

Kindleberger read aloud the resolution submitted by David M. Gordon and Robert B. Zevin:

> Let it hereby be resolved that: The Executive Committee of the American Economic Association, and the Finance Committee thereof, shall make a public declaration of intent not to buy any new securities, and to dispose within the next eighteen (18) months of all existing securities in companies either operating in or investing in the Republic of South Africa or in Namibia. The determination of companies either operating in or investing in the Republic of South Africa and Namibia shall be made with reference to lists maintained by the Investor Responsibility Research Center. The effect of this resolution shall be annulled if and only when a full elimination of the system of apartheid is achieved in the Republic of South Africa.

He then ruled the resolution out of order because it conflicted with the AEA's certificate of incorporation which provides, among other things, that "The Association as such will take no partisan attitude, nor will it commit its members to any position on practical economic questions." Kindleberger then recognized Gordon who moved to appeal the Chair's ruling; Zevin seconded. Gordon argued that the resolution was not out of order and cited Section IV, Article 6 of the bylaws. To wit, "The Executive Committee shall have control and management of the funds of the corporation. [It] may adopt any rules or regulations for the conduct of its business not inconsistent with the constitution or rules adopted at annual meetings." Zevin argued that the resolution did not involve a practical economic question since the action called for would not affect the risk-return of the portfolio; neither did the resolution involve a partisan attitude because apartheid was opposed by virtually all. The motion to appeal the out-of-order ruling of the Chair was passed, 55 to 53.

Rendigs Fels then moved, and it was seconded, to amend the resolution to read as follows:

> Let it hereby be resolved that: The Executive Committee of the American Economic Association and the Finance Committee thereof shall not buy any new securities and shall dispose within eighteen (18) months of all existing securities in companies either operating in or investing in the Republic of South Africa or in Namibia. The determination of companies either operating in or investing in the Republic of South Africa and Namibia shall be made with reference to lists maintained by the Investor Responsibility Research Center.

The amendment passed by voice vote. The resolution as amended then passed by a voice vote.

Kindleberger reminded the audience of the provision in the bylaws that allows the Executive Committee to submit resolutions adopted at an annual meeting in which less than 5 percent of the membership of the Association has voted to a mail ballot of the membership.

There being no additional business, Kindleberger introduced Alice Rivlin, President of the Association for 1986. The meeting was adjourned.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in New York, NY, March 22, 1985.

The first meeting of the 1985 Executive Committee was called to order at 10:10 a.m. on March 22, 1985 in the Rendezvous-Trianon Room of the New York Hilton. Members present were Charles P. Kindleberger (presiding), Orley Ashenfelter, Elizabeth E. Bailey, Alan Blinder, Rendigs Fels, Victor Fuchs, C. Elton Hinshaw, W. Arthur Lewis, Daniel McFadden, Janet L. Norwood, Alice M. Rivlin, Charles L. Schultze, A. Michael Spence, and Joseph Stiglitz. Leo Raskind, Counsel, was also present. Kindleberger welcomed Ashenfelter, Blinder, and Stiglitz to the Committee. Present for parts of the meeting were members of the Nominating Committee, Honors and Awards Committee, Robert McNeill, James Weiss, Donald Brown, and Marcus Alexis.

Minutes. A corrected set of minutes of the meeting of December 27, 1984 had been circulated prior to the meeting. They were approved without additional changes.

Investment Counselors. McNeill and Weiss, representing the Association's investment counselor, Stein Roe & Farnham, were present to discuss the management of the Association's portfolio as had been requested at the December 27, 1984 meeting. They reviewed the written report that had been circulated prior to the meeting (a copy is available from the Secretary) and responded to questions. It was decided that the portfolio probably did not take into sufficient account the tax-exempt status of the Association, and the Treasurer and Finance Committee were charged to consider the issue.

Report of the Secretary (Hinshaw). The Secretary reported that the 1985 annual meeting would be held in New York, December 28–30. The schedule for subsequent meetings is New Orleans (1986) and Chicago (1987). Atlanta, Boston, and New York are being considered as sites for 1988. Registration for the 1984 Dallas meetings totaled 5,065. Forty-one other associations, societies, and organizations met with the AEA, 364 scholarly sessions were held, and 96 other events (cocktail parties, breakfasts, committee meetings, lunches, etc.) were scheduled. The last time (1975) the ASSA met in Dallas, registration was 3,885, 23 other groups participated, 212 scholarly sessions were held, and 64 other events were scheduled. The 1984 meeting will yield a surplus of about $10,000.

A new directory of members will be published as a special December issue of the *American Economic Review*. Because of the peculiarities of regulations concerning the use of second-class postal privileges, it is cheaper to publish articles along with directory information than to publish only directory information. Since 1985 is the 100th birthday of the Association, five articles appropriate for the occasion and the directory have been commissioned: "Early American Leaders—The Neoclassical Tradition," James Tobin; "Early American Leaders—The Institutionalist and Critical Tradition," Martin Bronfenbrenner; "The Beginnings of Empirical Economics in America," Carl Christ; "Changes in Methods of Analysis and Research," William Baumol; and "The Expanding Domain of Economics," Jack Hirshleifer.

The Secretary had received three letters protesting job advertisements placed by King Saud University in the March 1984 issue of the *American Economic Review* because "It is well-known by all that this University will not employ an Economist belonging to the Jewish faith." Schultze received a similar letter from a representative of the Anti-Defamation League of B'nai B'rith protesting an ad by the University of Saudi Arabia. These same universities also list their job openings in *Job Openings for Economists*. Pending this meeting of the Executive Committee, the Secretary temporized by dividing the "Academic Listings" in *JOE* into two categories—United States and Foreign—and placed the following statement before the foreign job listings:

> These universities are not subject to U.S. laws and regulations concerning

equal opportunity employment. Some of their hiring procedures and employment practices may vary from those in the United States.

The Secretary asked for guidance about what policy the AEA should have concerning ads and job listings by foreign universities whose employment practices may be questionable by U.S. standards. The ensuing discussion raised several issues: Should the AEA take such charges at face value and act without specific evidence and an investigation? Could the AEA "police" the employment practices of advertisers? Would the universities mentioned above hire women? Could sectarian universities discriminate in favor of religious adherents? Was a "Surgeon General" type warning such as the Secretary placed in *JOE* sufficient? What responsibility does the AEA have to make known the availability of jobs in foreign countries to foreign graduate students in this country? It was decided to proceed in this specific case by directing the Secretary to write the universities in question, inform them of the allegations, and seek their response. Further action would depend on their answers.

A member had inquired about the possibility of the AEA sponsoring an excess major medical plan in addition to the life insurance plan. Although some concern was expressed about the Association becoming more involved in such business arrangements, the Secretary was asked to seek proposals for excess major medical insurance.

Although few may recall the event, the Regency Hotel of Denver sued the ASSA for breach of contract after the 1980 ASSA meetings in Denver. The court of original jurisdiction ruled in favor of the hotel. The ASSA appealed. The Colorado Court of Appeals has reversed the original decision and has ruled in ASSA's favor. Although the hotel can appeal to the Colorado Supreme Court, the ASSA lawyer, James Ruh, thinks it will not; even if it does, he doubts the Supreme Court would hear the case. For all practical purposes, the case is over.

Naomi Perlman filed a complaint with the Pennsylvania Human Relations Commission against the AEA alleging discrimination based on gender. Leo Raskind and the Secretary presented the AEA's side in a fact-finding hearing in Pittsburgh. Before the Commission made a determination, we filed a motion to dismiss. That motion was rejected. We have appealed. If the appeal is denied, the Commission will render a judgment.

Honors and Awards (Oliver Williamson). Acting together as an electoral college, the Committee on Honors and Awards and the Executive Committee voted to award the John Bates Clark Medal to Jerry A. Hausman.

Nominating Committee (Gardner Ackley). Acting together as an electoral college, the Nominating Committee and the Executive Committee chose Gary Becker as the nominee for President-elect and Joseph Pechman and Paul Rosenstein-Rodan as Distinguished Fellows. Ackley reported the following nominees for other offices: for Vice-President (two to be chosen), Richard Cooper, Peter Diamond, Mancur Olson, and Thomas Schelling; for members of the Executive Committee (two to be chosen), Donald Gordon, Sherwin Rosen, Thomas Sargent, and T. N. Srinivasan.

Foreign Honorary Members. Acting on the written report submitted by Richard Musgrave, Chair of the Committee on Foreign Honorary Members, the Executive Committee elected Anthony Atkinson as an honorary member. The Secretary was directed to write the Committee and ask them to consider additional nominations, especially economists from the "Third World."

1985 Program (Rivlin). In addition to the general program, which emphasizes the usefulness of economics for policy decisions, Rivlin is planning three special events to celebrate the centennial: a birthday party that features humorous skits lampooning economics; a special speaker at a luncheon on December 29; and a joint session with the History of Economics Society. The major portion of the program is pretty much set.

Committee on the Status of Minorities (Brown and Alexis). After hearing a review of the history of the program and discussing Brown's written proposal for a minority fellowship program in economics to be funded

by Rockefeller Foundation and the Federal Reserve System, the Executive Committee voted to accept responsibility for the administration and implementation of the two grants to establish the fellowship program as described in the proposal (a copy of which is available from the Secretary).

Search Committee for a Managing Editor of the Journal of Economic Literature (Ackley). It was voted to approve the Committee's recommendation of John Pencavel as Managing Editor of the *JEL* for a three-year term beginning January 1, 1986. It was understood that Abramovitz would become an Associate Editor and that Alexander Field would continue to handle the book reviews.

Report of the Managing Editor of the American Economic Review (Ashenfelter). Ashenfelter described the proposed structure of the editorial process. He sought approval for the appointment of three co-editors, each having final authority concerning papers sent to him by Ashenfelter for review. Submitted manuscripts would be divided into four, roughly equal, batches by substantive area and methodology used and referred to the appropriate editor. It was voted to approve the appointment of Robert Haveman, John Riley, and John Taylor as co-editors.

Ashenfelter went on to state that the Board of Editors would be expected to referee papers. Each of the members he was proposing (after consultation with the co-editors) understood that. It was voted to approve his recommendations: George Akerlof, Richard Schmalensee, Jacob Frenkel, Claudia Goldin, George E. Johnson, John Kennan, Mervyn King, Paul Krugman, Bennett McCallum, Edgar Olsen, Stephen Shavell, and John Shoven.

The backlog of manuscripts already accepted by Robert Clower will fill the *AER* through the March 1986 issue and possibly part of June. Should a temporary increase in size of each issue be allowed until the backlog is reduced? It was voted to authorize the President to decide the issue after more information on costs have been obtained from the Treasurer.

Report of the Managing Editor of the Journal of Economic Literature (Abramovitz). Abramovitz stated that the significant business concerning the journal had already been discussed when the appointment of John Pencavel to the editorship was considered and approved. He would only add that the possibility of adding a bibliography of working papers (developed and maintained by the University of Warwick) to *JEL*'s data base which is available on DIALOG is being considered.

Search Committee for an Editor of the New Journal (Schultze). The search is progressing, but slowly. He hopes to have a recommendation ready for the December 27th meeting.

Report of the Treasurer (Fels). In 1984 there was a small operating deficit more than offset by investment gains, resulting in an overall surplus of not quite a quarter of a million dollars, almost the same as 1983. A revision of the 1985 budget submitted to the Executive Committee in December shows a projected operating deficit of \$196 thousand partly offset by projected investment gains of \$147 thousand. The revised budget includes an increase for the *AER* of \$106 thousand. Part of this increase is temporary (estimated to be \$30 thousand), resulting from the move of the editorial offices. Part of it is permanent, resulting from the restructuring of the editorial function. It was voted to approve the revised budget.

Expenses are likely to continue to increase in 1986 and thereafter since the operations of the Association are labor intensive. Start-up costs of the new journal are likely to reduce investments by \$500 thousand, reducing future income perhaps \$20 thousand a year. Operating costs of the new journal will have to be financed for a time. Deficits are likely to begin in 1986 and continue unless dues and subscriptions are increased. It was voted to increase the base rate for dues from \$35 to \$37.50 and the subscription price from \$100 to \$105 effective January 1, 1986.

At the December 1984 meeting, the Executive Committee considered changing the investment income formula used since the 1960's to a simpler method—4 percent of the market value of the portfolio. No decision was made but the Treasurer was asked to compare the results of the two methods over several past years. He presented the comparison for the years 1969 to 1984. Total income

over the period was roughly the same; the income stream from the 4 percent formula was smoother.

Because of increasing fatigue, decreasing attendance, and the impending departure of planes and trains, no decision was made. The meeting adjourned at 4:45 P.M.

Minutes of the Meeting of the Executive Committee in New York, New York, December 27, 1985.

The second meeting of the 1985 Executive Committee was called to order at 10:00 A.M. on December 27, 1985 in the Rendezvous-Trianon Room of the New York Hilton Hotel, New York, New York. Members present were Charles P. Kindleberger (presiding), Moses Abramovitz, Orley Ashenfelter, Elizabeth E. Bailey, Alan S. Blinder, Rendigs Fels, Victor R. Fuchs, C. Elton Hinshaw, W. Arthur Lewis, Daniel McFadden, William D. Nordhaus, Janet L. Norwood, John Pencavel, Alice M. Rivlin, Charles L. Schultze, A. Michael Spence, and Joseph E. Stiglitz. Also present were Gary S. Becker, Mancur Olson (newly elected members of the 1986 Executive Committee), and Leo Raskind (Counsel). Donald J. Brown, Kalman Goldberg, Julie Marsh, Michael McCarthy, David Richardson, and Art Singer were present for parts of the meeting to present reports. President Kindleberger welcomed the new members and thanked those whose terms were expiring.

Minutes. The minutes of the previous meeting (March 22, 1985) were approved as written and circulated prior to this meeting.

Report of the Secretary (Hinshaw). The 1986 annual meeting of the Association will be held in New Orleans December 28–30. Chicago will be the site of the 1987 meetings. It was VOTED to approve the Secretary's recommendation of New York for 1988 and Atlanta for 1989 and authorize him to negotiate contracts with appropriate hotels in these cities.

At the last meeting, in response to complaints from members concerning discriminatory hiring practices by two Saudi universities, the Executive Committee directed the Secretary to write the universities, inform them of the allegations, and seek their response. Further action would depend on their response. The Secretary wrote but has received no answer. No job ads have been accepted pending a response to the inquiry. After a discussion about the need or desirability for a general policy statement concerning discriminatory hiring policies by some foreign universities, it was decided to ask Alan Blinder to draft a statement for consideration at the next meeting. Until then, the Secretary was to continue a case-by-case policy.

The Secretary reported on the best proposal he had received for an excess major medical insurance plan for members of the Association. It was decided that Charles Schultze would undertake a brief foray into the world of insurance and report his findings to the next meeting. The Executive Committee would then decide whether or not to proceed with developing such a plan.

Report of the Editor of the American Economic Review (Ashenfelter). Ashenfelter reviewed his written report (see elsewhere in this issue). The new editorial process decentralizes the decision-making procedure. Each manuscript submitted will be assigned to one of the three co-editors or retained by the managing editor. Robert Haveman, John Riley, and John Taylor are serving as co-editors. Taylor deals primarily with the macro area, Riley with industrial organization and microtheory, Haveman with public finance and public policy, and the managing editor handles the others. Each editor will be managing about 200 to 275 manuscripts throughout the editorial process each year at the current submission rate. Fewer manuscripts are being rejected without outside refereeing. The June 1986 issue will be the first that will contain articles accepted under Ashenfelter's editorship. It was VOTED to approve Ashenfelter's recommendation of Alvin Roth and Myron Scholes as new members of the *AER* Board of Editors.

Several members of the Association, in a letter to Kindleberger, had raised the issue of required disclosure of the sources of financial support for papers published in the journals of the Association and also of *ex parte* pro-

ceedings from which the work originated. Although the issue was raised because of an article accepted when Robert Clower was editor, Ashenfelter was aware of the situation and had discussed the matter with his co-editors. He planned to seek advice from his Board. During the discussion, it was noted that "disclosure" may not be enforceable, decisions about whether to reveal partial support or germs of ideas arising from consulting would not be easily made, and medical journals are moving toward a disclosure policy. The idea that "truth will out quickly" is probably not true; it may out eventually but it takes time. It seemed to be the consensus that some form of disclosure policy was desirable. A committee of the two Managing Editors and AEA Counsel was appointed to consider the issue and bring a recommendation to the next meeting of the Executive Committee.

Report of the Managing Editor of the Journal of Economic Literature (Abramovitz). Abramovitz reviewed his written report (see elsewhere in this issue). He pointed out that by 1987 the backlog of the annual *Index of Economic Articles* will have been eliminated; the log in publication will have been reduced as much as is technically possible—to about 18 months. Thereafter, one volume per year will be published. He also noted the increased development and use of DIALOG, the online computer access service to the subject and author indexes of the annual *Indexes*. The December 1985 issue of *JEL* contains an article by Bernard Saffran and Drucilla Ekwurzel on the use of the system, "Online Information Retrieval for Economists."

After his report, the Executive Committee gave Abramovitz a round of enthusiastic applause for the accomplishments during his term as editor. As of January 1, 1986, the editorship passes to John Pencavel. Pencavel stated that he expects to continue to pursue the general goals for the journal established by Abramovitz. There will not be a marked change in its tone and character.

The 1986 Program (Becker). President-elect Becker, program chair for 1986, said that he has appointed an advisory committee to help him with the program. There would be no general theme for all sessions, but 6 to 8 sessions would be devoted to the topic of the use of economics in new areas. The remainder of the program would cover the major areas.

COSSA and COPAFS. Written reports from the representatives to the Consortium of Social Science Associations (COSSA) and the Council of Professional Associations on Federal Statistics (COPAFS) had been circulated prior to the meeting. During the discussion of the reports, Rivlin stated that the AEA has been relatively passive in its attitude concerning the role of the federal government in data collection and support of basic research. Recent budget legislation suggests the possibility of a serious reduction in the quality of data. Norwood indicated that budget cuts would probably lead to the elimination of methodological descriptions of data sets. The quality of data will decline if not the quantity. Rivlin suggested that the Association may need to establish a standing committee to represent more effectively economists' interests in these areas. No action was taken.

Committee on Economic Education (Goldberg). On behalf of the Committee, Goldberg presented a request for funds to convene a symposium to explore various facets of the graduate education of economists. Most research and programs have emphasized precollege and undergraduate instruction. There has been little effort to examine graduate education. The topic has been neglected. The total cost of the proposed symposium was estimated to be $15,500.

Kindleberger noted that "Our blessing comes quickly, but $15,500 takes longer." It was suggested that the symposium be folded into the annual AEA meetings. If the proposed papers were published in the *Papers and Proceedings*, they would receive wider attention than if published elsewhere. It was also suggested that the Committee seek funds from foundations interested in the area. Some members of the Executive Committee expressed concern about funding a symposium; it could be a "slippery slope." The number of worthy symposia that might be

supported is large. It was VOTED not to fund the project. The Executive Committee agreed that the goals of the proposal were laudable but had reservations about the planned mechanism.

Committee on the Status of Women in the Economics Profession (Sawhill). Sawhill distributed a written report and briefly reviewed it with the Committee. The report is published elsewhere in this issue. It was VOTED to approve $15,000 plus a "cost-of-living" adjustment for the 1986 CSWEP budget and a $10,000 carryover from previous appropriations.

Committee on the Status of Minorities in the Economics Profession (Brown). Brown, Chair of the Committee, introduced four guests: David Richardson, Director of the summer program at the University of Wisconsin; Michael McCarthy, Director of the summer program when it moves to Temple University; Julie Marsh, Registrar of the Committee's fellowship programs; and Art Singer, representative of the Sloan Foundation. Brown reported on the 3 programs under the purview of his Committee—the summer program, the AEA minority fellowship program, and the AEA/Federal Reserve Bank fellowship program.

Thirty or so undergraduate minority students are selected from about 100 applicants to attend an eight-week instructional program in micro, macro, and math and statistics. Richardson reported on the program as it has operated at Wisconsin during the past three years. McCarthy reported on how the program will operate at Temple during the next three years. Singer reported that the Sloan Foundation has supported the program for fifteen years, but is phasing out its support. The Temple program will end its commitment. The Foundation is interested in a "follow-up" study of the impact of the program and would consider funding a proposal from the AEA to do one. The consensus of the Executive Committee was that such a study would be useful.

The AEA minority fellowship program is funded by the Rockefeller Foundation. These fellowships support minority students during their first two years in graduate school. Rockefeller support is currently scheduled to end in two years. If the program is to continue, new funds will have to be found.

The AEA/Federal Reserve fellowship program is funded by the Federal Reserve. It finances minority students during their third and fourth years of graduate work. It is contingent upon the continuation of funding for the first and second years. If funds for the first two years disappear, Federal Reserve funding will disappear.

In summary, during the next two years the AEA is faced with finding new funding for the summer program to replace the Sloan Foundation monies and for the minority fellowship program to replace the Rockefeller Foundation monies. It was VOTED to award the summer program to Temple University based on its proposal to the Sloan Foundation and to accept the Federal Reserve Bank Fellowship grant.

Foreign Honorary Members. Acting on the written report of the Committee on Foreign Honorary Members, it was VOTED to elect Michael Bruno, Max Cordon, and Frank Hahn as honorary members of the Association.

Honors and Awards. The Secretary was instructed to request the Committee on Honors and Awards to "cast a broader net" in seeking candidates for the Clark medalists. Specifically, department chairs and former Clark medalists should be solicited for nominations. Such nominations should contain careful evaluations of the candidates from specialists in their subjects, resumes, and statements of contributions.

New Journal (Schultze). After an extended discussion of the difficulties of finding or generating good expository articles for the proposed journal and other obstacles an editor would confront, it was VOTED to ask Joseph Stiglitz to present a detailed proposal, with a tentative budget for three issues, for the establishing and editing of the new journal. His proposal will be reviewed at the March 21, 1986 meeting of the Executive Committee.

Report of the Treasurer (Fels). The Treasurer reviewed his written report (see elsewhere in this issue). The finances of the Association are in good shape. A modest surplus is anticipated for 1985 and a small

surplus is budgeted for 1986. However, the 1986 budget makes no provision for the start-up of the new journal. It was VOTED to approve the proposed 1986 budget.

Other Business. The resolution requiring sale of stocks of companies doing business in South Africa which will be submitted to the general business meeting was discussed. No position was taken. It was understood that if the resolution passes in its original form, the Executive Committee would consider submitting it to a mail ballot of the membership.

There being no more business to conduct, the meeting adjourned.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1985

Annual Meetings. In 1986, the annual meeting will be held in New Orleans on December 28–30. The schedule for subsequent meetings is Chicago in 1987, New York in 1988, and Atlanta in 1989. Each of these meetings is scheduled for December 28–30 and each will have a Placement Service, which will open for business one day earlier (December 27) than the meetings.

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Gardner Ackley, Chair, Padma Desai, James A. Hefner, Ronald W. Jones, Stanley Lebergott, Martin F. Prachowny, and Thomas E. Weisskopf submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-Elect
Gary S. Becker

Vice-President	*Executive Committee*
Richard N. Cooper	Donald E. Gordon
Peter A. Diamond	Sherwin Rosen
Mancur Olson	Thomas Sargent
Thomas C. Schelling	T. N. Srinivasan

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots, I certify that the following persons have been duly elected to the respective offices:

President-Elect (for a term of one year)
Gary S. Becker

Vice-Presidents (for a term of one year)
Peter A. Diamond
Mancur Olson

TABLE 1—MEMBERS AND SUSCRIBERS
(End of Year)

	1983	1984	1985
Class of Membership			
Annual	16,728	16,612	17,602
Junior	1,998	1,932	1,670
Life	383	370	359
Honorary	31	29	30
Family	423	422	472
Complimentary	599	521	473
Total Members	20,162	19,886	20,606
Subscribers	5,986	5,846	5,852
Total Members and Subscribers	26,148	25,732	26,458

Executive Committee (for a term of three years)
Sherwin Rosen
Thomas Sargent

In addition, I have the following information:

Number of legal ballots	5,460
Number of invalid envelopes	216
Number of envelopes received after October 1	86
Number of envelopes returned	5,762

The 1985 Directory. The new directory of members was mailed in December. In addition to the usual biographical material, it contains five essays commemorating the centennial of the Association:

William J. Baumol, "On Method in U.S. Economics a Century Earlier"
Martin Brofenbrenner, "Early American Leaders—Institutional and Critical Traditions"
James Tobin, "Neoclassical Theory in America: J. B. Clark and Fisher"
Carl F. Christ, "Early Progress in Estimating Quantitative Economic Relationships in America"
Jack Hirshleifer, "The Expanding Domain of Economics"

The Association is fortunate to have Mary Winer as its Administrative Director. She planned, coordinated, and saw to completion the directory's publication.

Membership. The total number of members and subscribers is shown in Table 1. The total has fluctuated between 25,000 and 26,500 since 1975, when it reached an all time high of 26,787.

National Registry. The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 381 in 1985, compared to 265 in 1984. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

AEA Staff. Mary Winer, Kimberly Adair, Norma Ayres, Ersye Burns, Violet Sikes, Jacquelyn Woods, Laura Taylor, and Debra Juhasz handle the day-to-day operations of the Association. Barbara Weaver and Marlene Keefer organize the operation of the annual meeting. Their dedication and efficiency make the job of the Secretary tolerable. I wish to express my great gratitude for the excellent work they continue to do.

Committees and Representatives. Listed below are those who served the Association during 1985 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them all for their services.

Ad Hoc Committee to Select Articles on the History of the Association for the 1985 Directory
Moses Abramovitz, *Chair*
Charles P. Kindleberger
George J. Stigler

Budget Committee
Rendigs Fels, *Chair*
A. Michael Spence (1985)
Janet L. Norwood (1986)
Daniel McFadden (1987)
Charles P. Kindleberger, *ex officio*
Alice M. Rivlin, *ex officio*

Census Advisory Committee
Laurence Chimerene (1985)
Ronald L. Oaxaca (1985)
Joel Popkin (1985)
Richard E. Quandt (1985)
Ann D. Witte (1985)
Morris A. Adelman (1987)
Rosanne E. Cole (1987)
Ben E. Laden (1987)
Victor Zarnowitz (1987)

Committee on Economic Education
W. Lee Hansen, *Chair* (1985)
Kalman Goldberg (1985)
Campbell R. McConnell (1985)
Daniel H. Saks (1986)
Marianne A. Ferber (1986)
Michael K. Salemi (1987)
William B. Walstad (1987)
Rendigs Fels, *ex officio*

Economics Institute Policy and Advisory Board
Edwin S. Mills, *Chair* (1986)
W. Lee Hansen (1985)
John R. Moroney (1985)
Dwight Perkins (1987)
Lance E. Davis (1988)
Stefan H. Robock (1988)
Joseph Havlicek, Jr. (1989)
Teh-wei Hu (1989)

Finance Committee
Rendigs Fels, *Chair*
Robert Eisner (1985)

Robert J. Genetski (1986)
Robert G. Dederick (1987)

Committee on Honorary Members
Richard A. Musgrave, *Chair* (1986)
Hal R. Varian (1986)
Richard E. Caves (1988)
Franco Modigliani (1988)
J. Carter Murphy (1990)
Gordon C. Winston (1990)

Committee on Honors and Awards
Oliver E. Williamson, *Chair* (1987)
Robert Eisner (1987)
William Vickrey (1987)
Richard R. Nelson (1989)
Dale W. Jorgenson (1991)

Nominating Committee
Gardner Ackley, *Chair*
Padma Desai
James A. Hefner
Ronald W. Jones
Stanley Lebergott
Martin F. Prachowny
Thomas E. Weisskopf

Committee on Political Discrimination
Robert J. Lampman, *Chair* (1986)
Lester C. Thurow (1985)
Herbert Gintis (1986)
Richard R. Nelson (1986)
Benjamin J. Cohen (1987)
Clark W. Reynolds (1987)

Search Committee for Editor of Journal of Economic Literature
Gardner Ackley, *Chair*
Stanley W. Black
Alan S. Blinder
James Buchanan
Allen C. Kelley

Albert Rees
V. Kerry Smith

Committee on the Status of Minority Groups in the Economics Profession
Donald J. Brown, *Chair* (1985)
Bernard E. Anderson (1985)
Ronald L. Oaxaca (1985)
Samuel L. Myers, Jr. (1986)
Rhonda Williams (1986)

Committee on the Status of Women in the Economics Profession
Isabel V. Sawhill, *Chair* (1987)
Barbara R. Bergmann (1985)
Joseph A. Pechman (1985)
Cordelia W. Reimers (1985)
Aleta A. Styers (1985)
Lourdes Beneria (1986)
Bernadette Chachere (1986)
Mary Fish (1986)
Sharon B. Megdal (1986)
Michelle J. White (1986)
Karen Davis (1987)
Helen Junz (1987)
Joan J. Haworth
Charles P. Kindleberger, *ex officio*

AEA/SSRC Joint Committee on U.S.–China Exchanges
Gregory C. Chow, *Co-Chair*
Kenneth Arrow
Lawrence R. Klein
Theodore W. Schultz

Committee on U.S.–Soviet Exchanges
Franklyn D. Holzman, *Chair* (1987)
Jennifer R. Reinganum (1986)
Lloyd G. Reynolds (1986)
Abram Bergson (1987)
Joseph A. Pechman (1988)
Richard N. Rosett (1988)

COUNCIL AND OTHER REPRESENTATIVES

American Association for the Advancement of Science, Section K, Social, Economic and Political Sciences
Roger Bolton (1985)

American Association for the Advancement of Slavic Studies
Judith Thornton (1985)

American Council of Learned Societies
C. Elton Hinshaw

Review Board of the American Statistical Association-Bureau of Census Fellowships
Zvi Griliches

Consortium of Social Science Associations (COSSA)
Henry J. Aaron
C. Elton Hinshaw

Council of Professional Associations on Federal Statistics (COPAFS)
George Jaszi
Walter S. Salant

Federal Statistics Users Conference
Paul Wonnacott (1985)

Internal Revenue Service Conference—Tax Administrative Research Strategies
Harvey Galper (1985)

International Economic Association
Kenneth Arrow
C. Elton Hinshaw

Policy Board of the Journal of Consumer Research
Louis L. Wilde (1985)

National Bureau of Economic Research
David A. Kendrick (1987)

Social Science Research Council
Hugh T. Patrick (1987)

U.S. National Commission for UNESCO
Walter S. Salant (1985)

REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1985

Inaugurations
Richard P. Triana, Clark University
Hilda Kahne
James Edmund Halligan, New Mexico State University
Nathaniel Wollman
Jack Wood Humphries, Sul Ross State University
Robert D. Tollen
William H. Likins, Greensboro College
Basil G. Coley
Herb F. Reinhard, Jr., Morehead State University
Glenn C. Blomquist
Arnold R. Weber, Northwestern University
Sidney Davidson
Michael J. Adanti, Southern Connecticut State University
Peter M. Costello
William Van Muse, University of Akron
Douglas Stewart
Daniel Berg, Rensselaer Polytechnic Institute
James P. Moran
Paul R. Verkuil, The College of William and Mary
David A. Whitaker
Wallas K. Beal, Central Connecticut State University
Paul L. Altieri
George Rupp, William Marsh Rice University
Patricia N. Pando
John J. Casteen, III, The University of Connecticut
Peter S. Barth
Frank Horton, Oklahoma University
Frank G. Steindl
Dorothy Ingling MacConkey, Davis and Elkins College
Lewis Bell

ASSA 1985 CONVENTION COMMITTEE

Peter Fousek, *Chair*
Barbara D. Russell, *Vice Chair*
Barbara Weaver, *Convention Manager*
Norma J. Ayres
Jean-Ellen Giblin
Alice Christensen
Lois Banks
Joann Martens
Ellen Trust
Arlene Eskin
Violet Sikes
Barton Sotnick
Benjamin A. Michalik
Marilyn Rubin
Janet Aschenbrenner
Marlene Keefer

C. ELTON HINSHAW, *Secretary*

Report of the Treasurer for the Year Ending December 31,1985

As the accompanying table shows, the finances of the American Economic Association are in good shape. There were large surpluses in 1982, 1983, and 1984. At this time, audited results for 1985 are not available, but another surplus is anticipated. Investment income more than offsets the operating deficits that began in 1984 and are continuing.

The budget for 1986 shown in the table was approved by the Executive Committee at its meeting on December 27, 1985. It makes no provision for the new journal the Executive Committee has decided to start. If money is spent on it in 1986, the surplus will be smaller than the table indicates.

The net worth of the Association is now about one and one-half times annual expenditures, more than enough for safety. Although the Association can easily afford to spend half a million dollars on startup costs for the new journal, doing so will reduce

TABLE —1986 BUDGET, AMERICAN ASSOCIATION
(thousands of dollars)

	First Nine Months (Unaudited)		Full Year Actual	Full Year Budgeted	
	1984	1985	1984	1985	1986
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	$580	$614	$782	$782	$850
Nonmember subscriptions	452	456	608	600	638
Subtotal	1,033	1,070	1,390	1,382	1,488
Subscriptions, *Job Openings for Economists*	18	20	28	28	28
Advertising	72	79	102	102	105
Sale of *Index of Economic Articles*	7	47	12	100	100
Sales of copies, republications, handbooks	21	22	28	28	32
Sale of mailing list	21	26	39	40	50
Annual meeting	21	16	21	5	16
Sundry	36	45	53	53	60
Total Operating Revenue	**1,229**	**1,323**	**1,674**	**1,738**	**1,879**
PUBLICATION EXPENSES					
American Economic Review	376	481	476	582	587
Journal of Economic Literature	529	569	716	747	758
Directory	49	41	65	68	70
Job Openings for Economists	35	35	50	55	55
Index of Economic Articles	4	41	10	72	100
Subtotal	993	1,167	1,317	1,524	1,570
OPERATING AND ADMINISTRATIVE EXPENSES					
General and Administrative	190	187	292	310	320
Committees	35	16	54	50	50
Support of other organizations	48	48	50	50	56
Subtotal	274	251	395	410	426
Total Expenses	**1,267**	**1,418**	**1,712**	**1,934**	**1,996**
OPERATING GAIN (LOSS)	(38)	(95)	(38)	(196)	(117)
INVESTMENT GAIN (LOSS)	199	171	285	147	250
SURPLUS (DEFICIT)	$ 162	$ 76	$ 247	$ (49)	$ 133
Ratio, Net Worth to Annual Expenses			**1.39**	**1.48**	

Report of the Managing Editor

American Economic Review

This report covers operations of the editorial office of the *American Economic Review* while it was at UCLA and, since the spring of 1985, while it has been at Princeton. Most of the statistical material in this report inevitably covers operations and decisions of the office while it was at UCLA and while Robert Clower was the managing editor. As indicated by Clower in his report of last year, the first papers that will be published based on submissions to the Princeton office will appear in the June 1986 issue of the *Review*.

A new system that significantly decentralizes the editorial decision-making process through the use of co-editors has been implemented. Experience with the new process is still limited, however, and a more complete report on the results of its implementation will have to await another occasion.

Manuscripts submitted for publication in the *Review* since March 1, 1985 have been handled by a new procedure. Each paper was assigned to one of three co-editors, or retained by the managing editor, for supervision through the review process and for a final decision on the manuscript status. Robert H. Haveman, University of Wisconsin-Madison, John G. Riley, University of California-Los Angeles, and John Taylor, Stanford University, have served as co-editors in this new endeavor.

It is my hope that our new editorial procedure will allow due recognition in the editorial process for the wide variety in both the substantive and methodological character of manuscripts submitted to the *Review*. It should also be possible to obtain these benefits without a significant increase in the length of time required to review and process submitted manuscripts.

I am especially pleased to report that Wilma St. John, production editor of the

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1966–85

Year	Submitted	Published	Ratio Published-to-Submitted
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15
1985	952	128	.13

TABLE 2—SUMMARY OF CONTENTS, 1984 AND 1985

	1984		1985	
	Number	Pages	Number	Pages
Articles	52	721	54	803
Shorter Papers, including Comments and Replies	86	358	74	373
Dissertations		22		22
Announcements and Notes Section		71		48
Index		10		10
Total		1182		1256

Review for eighteen years, is continuing to serve in that capacity at the new editorial office in Princeton.

The history of the 1985 submissions and papers published is shown in Tables 1 and 2. We received 952 papers, an increase over 1984 of 31 submissions.

The subject matter distribution of papers published in 1984 and 1985 is shown in Table 3. It is my impression (as it has been with past managing editors) that the distribution of published papers reflects fairly accurately the distribution of papers submitted.

TABLE 3—SUBJECT MATTER DISTRIBUTION OF PUBLISHED MANUSCRIPTS, 1984 AND 1985

	Published	
	1984	1985
General Economics and General Equilibrium Theory	14	18
Microeconomic Theory	13	12
Macroeconomic Theory	7	15
Welfare Theory and Social Choice	10	8
Economic History, History of Thought, Methodology	2	5
Economic Systems	2	5
Economic Growth, Development, Planning, Fluctuations	2	2
Economic Statistics and Quantitative Methods	9	5
Monetary and Financial Theory and Institutions	6	10
Fiscal Policy and Public Finance	8	5
International Economics	10	5
Administration, Business Finance	7	2
Industrial Organization	11	10
Agriculture, Natural Resources	8	3
Manpower, Labor Population	13	15
Welfare Programs, Consumer Economics, Urban and Regional Economics	8	8
Total	138	128

Table 4 shows the printing and mailing expenses for the four regular issues and for the *Papers and Proceedings* issue of the *Review* for 1985. As in earlier years, the *Papers and Proceedings* continue to account for approximately 25 percent of total printing and mailing expenses.

The eighth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1985. As in 1984, this task was handled by John Riley, associate editor at that time, and by Wilma St. John. As in past years, manuscripts were processed directly to pageproofs.

The Board of Editors now consists of eighteen members, chosen by the managing editor in consultation with the co-editors and with approval of the Executive Committee of the Association. The Board has been helpful in refereeing, in dealing with comments on published articles, and in reading papers that are the subject of complaint over the fairness or competence of referees. I am grateful to the members of the Board for their assistance and advice, for their generally supportive attitude, and for useful criticism of particular actions of the managing editor, or particular aspects of the operations of the *Review*.

Two members of the Board would have completed their terms as of March 31, 1985 (as reported last year), but graciously agreed to extend their terms for another year: George Akerlof and Richard Schmalensee. I

TABLE 4—COPIES PRINTED, SIZE, AND COST OF MAILING, 1985 *AER*

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	292	336	$53,257.27	$1,052.86	$54,310
May	28,000	463	488	77,307.29	3,799.61	81,107
June	27,500	304	336	54,336.53	1,416.08	55,752
September	27,500	326	368	60,670.20	1,610.93	62,281
December[a]	28,000	334	384	62,550.00	1,200.00	63,750
Annual Misc.[b]						7,800
Total		1,719	1,912	$308,121.29	$9,079.48	$317,200

[a] Estimated.
[b] Estimated—Based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

am most grateful to them, and to the continuing members: Clive Bull, Michael R. Darby, Philip E. Graves, Meir Kohn, Susan Woodward, and Leslie Young. I wish to thank also the new members of the Board who were approved for their three-year terms that began April 1985: Jacob A. Frenkel, Claudia D. Goldin, George E. Johnson, John F. Kennan, Mervyn A. King, Paul Krugman, Bennett T. McCallum, Edgar Olson, Steven Shavell, and John B. Shoven.

I wish to thank the outgoing staff at UCLA, Theresa De Maria and Marcus Hennessey, for dedicated performance and assistance in closing that office. Our editorial assistants at the Princeton office, Shirley J. Griesbaum and Sandra D. R. Grant, are also owed a special debt of gratitude for helping to open up the new office. I also wish to express my gratitude to our graduate mathematics consultants Fred Luk and Mario Rui Pascoa at UCLA, and to Ann Case and Lorin Kusmin at Princeton.

We list below the referees used by Robert Clower in 1985 and in the last quarter 1984, and add our thanks to his for their time and energy devoted to the advancement of our science.

M. Abbott
A. B. Abel
J. M. Abowd
K. Abraham
I. Adelman
J. Aizenman
G. Akerlof
M. A. Akhtar
A. Alchian
J. G. Altonji
E. Applebaum
C. Archibald
K. Arrow
R. Ayanian
C. Azariadis
D. Backus
M. N. Baily
R. E. Baldwin
E. Baltensperger
D. P. Baron
R. J. Barro
F. M. Bass
W. J. Baumol
G. S. Becker
J. R. Behrman
L. Benham
E. Berglas
B. Bernanke
E. R. Berndt
T. J. Bertrand
H. Bester
J. Bhagwati
O. J. Blanchard
M. Blaug
A. S. Blinder
R. W. Boadway
M. D. Bordo
G. Borjas
H. Bowen
S. Bowles
D. Bradford
W. C. Brainard
J. A. Brander
M. Bray
D. Breeden
M. J. Brennan
T. F. Bresnahan
D. Brookshire
C. C. Brown
J. Brown
E. K. Browning
J. Brueckner
J. Bryant
W. H. Buiter
C. Bull
J. Bulow
J. B. Burbidge
R. Burkhauser
P. Cagan
G. Calvo
D. R. Capozza
D. Card
J. Carmichael
L. Carmichael
C. Chamley
G. C. Chow
C. Christ
K. Clark
P. K. Clark
C. Clotfelter
A. W. Coats
R. A. Cohn
J. Conlisk
R. Cooper
B. Cornell
V. Crawford
M. L. Cropper
A. Cukierman
J.-P. Danthine
M. Darby
J. DaVanzo
L. DeAlessi
F. Deleeuw
H. Demsetz
A. Denzau
W. G. Dewald
D. N. Dewees
C. Diaz-Alejandro
D. Dollar
R. Dornbush
G. M. Duncan
B. C. Eaton
A. Edwards
S. Edwards
I. Ehrlich
R. Eisner
B. C. Ellickson
L. G. Epstein
R. Evenson
R. C. Fair
R. Faith
R. E. Falvey
E. F. Fama
H. S. Farber
R. Farmer
W. Fellner
G. Fethke
G. S. Fields
F. M. Fisher
R. H. Frank
J. Frankel
J. A. Frenkel
D. Friedman
D. Fullerton
F. Gahvari
D. C. Gale
N. Gallini
J. A. Gapinski
P. M. Garber
G. Garvey
J. Gerson
M. Gersovitz
F. Gollop
R. J. Gordon
E. Gramlich
P. Graves
J. Green
J. M. Griffin
G. M. Grossman
H. I. Grossman
J. D. Gwartney
R. Hall
J. Haltiwanger
J. Ham
D. S. Hamermesh
B. Hamilton
M. Harris
J. C. Harsanyi
R. C. Hartman
J. Hasbrouck
M. Hashimoto
R. H. Haveman
R. Heiner
R. Heinkel
D. F. Hendry
A. O. Hirschman
D. Hirshleifer
J. Hirshleifer

R. J. Hodrick
W. Holahan
S. Hollander
C. A. Holt
P. Howitt
D. Hsieh
J. Huizinga
C. R. Hulten
G. Johnson
J. Jondrow
R. Jones
B. Jovanovic
J. Judd
J. P. Kalt
C. Kahn
M. Kamien
J. Kareken
M. L. Katz
J. Kennan
N. Kiefer
R. Kihlstrom
M. Killingsworth
K. Kimbrough
A. Kleidon
M. Kohn
J. Kornai
L. Kotlikoff
K. Krishna
C. Krouse
D. Laidler
J. Laitner
G. H. Lamson
E. M. Landes
K. Lang
V. LaVia
E. Leamer
A. Leijonhufvud
J. S. Leonard
S. F. LeRoy
D. Levine
S. J. Liebowitz
C. M. Lindsay
A. N. Link
R. E. Lucas, Jr.
R. F. Lucas
S. Lundberg
J. J. McCall
R. E. McCormick
M. J. McKelvey
W. McManus
J. McMillan
L. J. Maccini
T. Macurdy
J. H. Makin
E. Malinvaud
N. G. Mankiw
J. Marchand
R. Masson
T. Mayer
W. Meckling
J. L. Medoff
R. O. Mendelsohn
J. Mincer
L. Mirman
H. Mohring
M. Montgomery
D. Mortensen
G. Mosseti
J. Muellbauer
J. F. Muth
M. C. Nerlove
D. M. G. Newbery
J. Newhouse
D. C. North
W. Novshek
R. Oaxaca
E. O. Olsen
M. Olson, Jr.
J. Ostroy
A. J. Oswald
J. C. Panzar
M. Parkin
D. O. Parsons
P. Pashigian
J. Peek
S. Peltzman
J. H. Pencaval
M. H. Pesaran
E. S. Phelps
R. S. Pindyck
C. Pissarides
M. Plant
C. Plott
I. P'ng
W. Poole
J. Quinn
R. Radner
J. Raisian
M. Ransom
A. Raviv
M. Reder
C. E. Reid
J. Reinganum
R. Rob
A. J. Robson
W. P. Rogerson
T. Romer
H. S. Rosen
M. Rothschild
D. L. Rubinfield
A. Rubinstein
T. Russell
R. Sah
D. Sappington
T. J. Sargent
T. R. Saving
J. L. Scadding
D. T. Scheffman
T. Schelling
F. M. Scherer
R. Schmalensee
A. J. Schwartz
M. Schwartz
G. W. Schwert
U. Segal
C. Shapiro
D. Shapiro
W. Shughart II
J. J. Siegel
C. A. Sims
A. Skinner
D. Slottje
M. Smirlock
J. Smith
K. V. Smith
R. S. Smith
V. L. Smith
A. Snow
K. Sokoloff
G. R. Solon
R. M. Solow
R. Startz
G. J. Stigler
J. E. Stiglitz
R. Stillman
A. C. Stockman
C. Stuart
J. Sweeney
L. G. Telser
P. Temin
R. Thaler
J. Tirole
S. Titman
R. D. Tollison
R. Topel
R. Tresch
L. Tyson
D. Ulph
R. Verrecchia
W. K. Viscusi
M. Waldman
N. Wallace
B. Weingast
R. Weintraub
L. Weiss
J. F. Weston
J. Whalley
L. H. White
W. White
J. Wilcos
J. G. Williamson
S. G. Winter
R. Wintrobe
D. A. Wise
G. Woglom
K. Wolpin
S. Woodward
F. C. Wykoff
L. Young

ORLEY ASHENFELTER, *Managing Editor*

Report of the Managing Editor

Journal of Economic Literature

During 1985, including the December issue soon to be distributed, the *Journal* published nine major articles, two review articles, and 160 book reviews. The articles in 277 journals were indexed, and those from 79 were abstracted. Some 300 pages were needed to carry the annotated list of new books.

Since 1981, when the editorship passed from Mark Perlman to me, the total number of pages in the *Journal* has grown by about 12 percent. The rise has, indeed, been concentrated in the last two years. It is due almost entirely to an enlargement of the *Journal*'s bibliographical and indexing departments. The Contents of Current Periodicals, Subject Index and Author Index together, grew from 1,081 to 1,282 pages since 1983, an increase of over 18 percent.

The enlargement of the *Journal*'s bibliographic work mainly reflects the continuing increase in the number of economics journals in this country and around the world. This growth poses a problem for the *Journal*—how to make its indexes and abstracts as helpful as possible to teachers, students, and researchers, while limiting the increase of the *Journal*'s size and expense. In 1984, the *Journal* staff, with the help of panels of specialists in different fields, made a survey of English language journals, which constitute a large majority of the journals on the *JEL* list. Our aim was to identify those older journals, whose value to economists had become marginal and those newer journals of greater value that should be added to our list. As a result of the survey, the *Journal*'s Board approved the de-listing of 38 journals, and these were dropped in the course of 1985. At the same time, 29 journals, which were either recommended by our panels or were new applicants, were added. This net reduction in the number of journals limited the expansion of our indexes. Yet the indexes did expand because the journals added carried more articles of interest to economists than those dropped. Presumably our indexes are now better suited to the needs of economists than they had been.

A similar survey of foreign language journals was carried out during the past year with the help of colleagues abroad. We expect to revise our list of journals accordingly, dropping some and adding others recommended by our panel of consultants. We also made a mail survey of the AEA membership to find out whether members thought it useful to continue printing the tables of contents of those foreign journals that do not provide English summaries of their articles. A majority of the respondents favor continuing this practice, and the staff will so recommend to the Board.

Our effort to reduce the lag in the publication of the annual *Index of Economic Articles* is showing results. Classification of articles by subject and author for the 1980, 1983, and 1981 volumes was completed during the year. The 1980 *Index* was published; 1983 is in press and will be published early in 1986. The 1981 volume is scheduled for publication in the fall 1986; and 1982 will follow. If, as we expect, 1984 and 1985 appear during 1987, the lag will have been reduced as much as is technically possible. We can then return to publishing one volume per year. It will appear approximately 18 months after the publication of the articles contained in its indexes.

The subject and author indexes of journal articles from 1969 through the current issue are now available by online computer access in the *Economic Literature Index* (*ELI*) of the DIALOG Information Retrieval Service. (See the Note published at the end of the *JEL* table of contents, each issue.) *ELI* also carries the articles that appeared in collective volumes from 1969 to 1979 (1979 was added in 1985), and 1980, 1981, and 1983 will be added in 1986.

The *Journal* staff has been active in providing instruction in the use of *ELI* on DIALOG. The December 1985 issue of *JEL*

carries an article by Bernard Saffran and *Journal* Associate Editor, Drucilla Ekwurzel, "Online Information Retrieval for Economists." Mrs. Ekwurzel also conducted instructional sessions in two DIALOG-sponsored workshops during the past year. The staff will again illustrate *ELI* searches at the *JEL* exhibit booth during the December AEA meeting.

In its continuing effort to control costs, the Pittsburgh office has installed a computer system. It will now be possible to eliminate certain steps in the composition and printing operations, and to achieve a measure of cost reduction in 1986 and thereafter. The computer will also facilitate the transmission of information to the DIALOG database.

I should like to express my warm thanks to the *Journal*'s staff both in Pittsburgh and Stanford. My special thanks go to Lyndis Rankin who retires at the end of 1985 after many years of highly effective work. She had an important part in the development of the *Journal*'s bibliographic data. She has earned the gratitude and good wishes of all economists.

The continuing full-time members of the staff at Stanford were Ann Vollmer and Anita Makler; at Pittsburgh, they were Patricia Andrews and Elizabeth Thornton. Beginning October 1, Linda Scott joined the Pittsburgh staff as Assistant Editor.

I should also like to express my thanks to the Board of Editors, our Consulting Editors abroad, and the many referees who have helped develop the *Journal*'s articles.

During 1985, the editorial duties of the *Journal* were shared by its Associate Editors: John Pencavel, who had particular responsibility for the Stanford office; Alexander Field, who is in charge of the Book Review department; Drucilla Ekwurzel, who has charge of the bibliographic work in Pittsburgh, and Editorial Consultant Asatoshi Maeshiro, who is responsible for classification and indexing. I am grateful to all of them for their help during my term as editor.

As of January 1, 1986, the Managing Editorship of the *Journal* passes to John Pencavel. It goes with my warm good wishes and full confidence. I continue on the staff to help in the Articles department and in liaison between the Pittsburgh and Stanford offices.

MOSES ABRAMOVITZ, *Managing Editor*

Report of the Director

Job Openings for Economists

The number of new jobs listed this year decreased by 7 percent from last year. In 1984, 1,713 new vacancies were advertised; this year 1,592 new jobs were listed. Both academic and nonacademic listings decreased. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type (academic or nonacademic) for each issue of *JOE* in 1985.

Universities with graduate programs and four-year colleges continue to be the major sources of job listings. Together they constitute about 80 percent of total employers. Table 2 shows the number of employers by type for each 1985 issue.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in

TABLE 1—JOB LISTINGS FOR 1985

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	84	140	67	106
April	44	64	41	58
June	29	49	26	45
August	34	74	30	70
October	157	403	137	371
November	140	289	140	289
December	206	459	105	224
Subtotal	694	1,478	546	1,163
Nonacademic				
February	17	58	14	44
April	15	45	13	35
June	17	54	15	44
August	16	62	13	42
October	33	132	30	117
November	20	45	20	45
December	42	201	19	102
Subtotal	160	597	124	429
TOTAL	854	2,075	670	1,592

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1985

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	42	42	3	–	2	1	8	3	101
April	23	21	3	2	1	–	7	2	59
June	11	18	2	2	3	2	5	3	46
August	17	17	2	–	4	3	5	2	50
October	60	97	11	1	10	–	9	2	190
November	51	89	8	2	3	–	5	2	160
December	81	125	15	–	8	1	16	2	248
TOTAL	285	409	44	7	31	7	55	16	854

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1985

Fields[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	74	43	25	47	189	154	229	761
Growth and Development (100)	14	13	10	10	42	18	42	149
Economics and Statistics (200)	32	16	15	19	72	42	96	292
Monetary and Fiscal (300)	32	13	8	19	97	60	122	351
International Economics (400)	18	10	7	11	51	46	67	210
Business Administration, Finance, Marketing and Accounting (500)	25	13	14	10	53	33	62	210
Industrial Organization (600)	24	18	19	22	62	33	56	234
Agriculture and Natural Resources (700)	10	14	9	9	26	10	32	110
Labor (800)	11	6	6	14	39	31	46	153
Welfare and Urban (900)	16	13	4	7	53	25	73	191
Related Disciplines (A00)	4	3	2	–	7	8	16	40
Administrative Positions (B00)	7	4	3	3	11	12	13	53
TOTAL	267	166	122	171	702	472	854	2,754

[a] Fields of specialization codes are from the *Journal of Economic Literature*.

mathematics and statistics appear to be the type of economist that employers are seeking. The applied area of specialization seems to be of secondary importance. Table 3 shows the number of citations by field of specialization. General economic theory (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200). This pattern has prevailed for the past several years.

Violet Sikes is almost solely responsible for the publication and distribution of *JOE*. I wish to express my great gratitude for the excellent job she continues to do. The Association is fortunate to have her exemplary services.

C. ELTON HINSHAW, *Director*

Policy and Advisory Board of the Economics Institute

The year 1985 was a year of slow recovery for the Economics Institute, as it was for the developing countries, from which most of the Institute's students and revenues come.

In 1985, 436 students attended the Institute. These students came from 56 countries and enrolled in 77 U.S. universities. As in previous years, Institute students are divided among graduate programs in economics, business, and agricultural economics. In addition, to its primary task of preparing students for U.S. graduate programs, the Institute also conducts short-term training programs for foreign professionals who return to posts in their home countries upon completion of the program.

For some years, the Institute has been mostly financed by tuition and fees paid by organizations in countries that send students. Some students are financed by assistance from U.S. sources. In addition, the Institute has a tiny scholarship fund of its own. The Institute's greatest need is for additional scholarship funds to finance students, mostly from South Asia and Africa, who cannot be financed from their home countries.

The Board held its annual meeting November 15 and 16, 1985, in Boulder, Colorado. In addition, several Board members visited the Institute during the summer of 1985, and several met at the ASSA meetings in New York.

Joseph Havlicek and Teh-wei Hu joined the Board. W. Lee Hansen agreed to a two-year extension of his term on the Board.

WYN F. OWEN, *Director*

EDWIN S. MILLS, *Chairman*

Report of the Representative to the International Economic Association

The Executive Committee will (in principle) recall that I was appointed representative of the American Economic Association to the International Economic Association in March 1983, for a three-year term. I was elected President of the International Economic Association in September 1983 by its Council, at a meeting in San Sebastian, Spain. The President's term is three years.

The Council of the IEA meets triennially, in conjunction with a World Congress. The Executive Committee of the IEA meets annually. The main functions of the IEA are the holding of conferences, two or three a year, and the triennial congresses. The *Proceedings* of the conferences and congresses are published by Macmillan (U.K.). Recent conferences are as follows: "Economic Incentives," organized by Herbert Giersch (FRG) and Bela Belassa (USA), June 1984, in Kiel (FRG); "East-West Economic Relations in the Changing Global Environment," Budapest (Hungary) and Vienna (Austria), October 1984, organized by Bela Csikos-Nagy (Hungary) and Friederich Levcik (Austria); and, "Peace, Defence, and Economic Analysis," organized by Frank Blackaby (Stockholm Institute for Peace Research) and Christian Schmidt (France), October 1985. A planned conference on, "Income Policies," to have been held in early September 1985, in Mexico City and organized by Victor Urquidi (Mexico), was cancelled on less than a month's notice because of Mexican financial difficulties. (Each conference is financed by local public and private sources.) For 1986, there are planned two more conferences, and the Eighth World Congress, to be held in New Delhi, India, December 1–5. We are indebted to the generosity of the Indian government and private banks. About $500,000 has been appropriated for this purpose.

The present Executive Committee has nominated the officers and Executive Committee members for 1986–89; the list has to be acted on by the Council at its meeting with the Congress in New Delhi. Amartya K. Sen, of Oxford, has been nominated for President, Bela Csikos-Nagy (Hungary) for Vice-President, and Luis Angel Rojo (Bank of Spain) for a second term as Treasurer.

KENNETH J. ARROW, *Representative*

Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research conducts analyses on a large variety of economic issues; publishes books, working papers, and two periodicals; sponsors conferences; and holds workshops and seminars as part of an annual summer institute. Approximately 280 economists at universities across the United States contribute to NBER's working paper series, and many other economists, here and abroad, attend conferences and the summer institute.

Programs. NBER's research is organized into eight programs (directors in parentheses): Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity and Technical Change (Zvi Griliches). Program meetings are generally held twice during the academic year and once, for a longer period, during the summer institute. About 250 people attended summer institute meetings in Cambridge in 1985.

Projects. The NBER also sponsors large-scale projects which bring together researchers from several of these programs. One major project centers on the role of Government Budget and the Private Economy. It includes the following subprojects (directors in parentheses): the impact of taxation on such behavior as charitable contributions (Charles Clotfelter); measuring and analyzing the role of state and local government in the economy (Harvey Rosen); studies of the compensation of public sector employees (David Wise); the impact of public sector unionization (Richard Freeman); and an analysis of government debt and deficits and their impact on the private sector (David Bradford and Benjamin Friedman).

A second major project, Productivity and Industrial Change in the World Economy, likewise has several major parts. William Branson and J. David Richardson are directing a project on international economic policy. Research on trade relations and trade policy is directed by Robert Baldwin. Richard Marston leads a group studying the effects of misaligned exchange rates.

In addition, NBER is currently sponsoring several smaller projects. Jeffrey Sachs is directing a project on Developing Country Debt. Richard Freeman heads a study of International Migration. Alan Auerbach is leading an examination of Mergers and Acquisitions. Martin Feldstein is directing a study of the Effects of Taxation on Capital Formation. And David Wise heads a project on the Economics of Aging.

Conferences. In 1985, NBER sponsored conferences (organizers in parentheses) in the United States and abroad on the following topics: Pensions in the U.S. Economy (John Shoven); Trade Policy in the Second Reagan Administration (Robert Baldwin); Money and Financial Markets (Robert Shiller); International Seminar on Macroeconomics (Robert Gordon and George de Menil); Economics Fluctuations (Robert Hall); Current Issues in U.S. Trade Policy (Robert Baldwin and J. David Richardson); Productivity Growth in the U.S. and Japan (Charles Hulten and Randall Norsworthy); Current Policy Issues in the U.S. and Japan (Geoffrey Carliner); and International Aspects of Fiscal Policies (Jacob Frenkel).

Books. In 1985, the following NBER books were published by the University of Chicago Press: *Social Experimentation* (Jerry A. Hausman and David A. Wise, eds.); *Corporate Capital Structures in the United States* (Benjamin M. Friedman, ed.); *Federal Tax Policy and Charitable Giving* (Charles T. Clotfelter); *A General Equilibrium Model for Tax Policy Evaluation* (Charles L. Ballard, Don Fullerton, John B. Shoven, and John B. Whalley); *Pensions, Labor, and Individual Choice* (David A. Wise, ed.); *Horizontal Equity, Uncertainty, and Measures of Well-Being* (Martin David and Timothy Smeed-

ings, eds.). In addition, Ballinger Publishing Company published *Monitoring Business Cycles in Market-Oriented Countries: Developing and Using International Economic Indicators* (Philip A. Klein and Geoffrey H. Moore, eds.).

In 1986, the following NBER books will be published: *Financing Corporate Capital Formation* (Benjamin Friedman, ed.); *The Black Youth Employment Crisis* (Richard B. Freeman and Harry Holzer, eds.); *Studies in State and Local Public Finance* (Harvey S. Rosen, ed.); *The American Business Cycle: Continuity and Change* (Robert J. Gordon, ed.); *Economic Adjustment and Exchange Rates in Developing Countries* (Sebastian Edwards and Liaquat Ahamed, eds.); *Long-Term Factors in American Economic Growth* (Stanley L. Engerman and Robert E. Gallman, eds.); *Issues in Pension Economics* (Zvi Bodie, John B. Shoven, and David A. Wise, eds.).

Periodicals. NBER publishes two periodicals, the *Digest* and the *Reporter*. The *Digest* provides summaries each month on recent NBER working papers of general interest. The quarterly *Reporter* contains longer summaries of recent program activity, reports of NBER conferences, reviews of recent work by NBER researchers, and abstracts of working papers issued during the previous quarter.

During 1985 Martin Feldstein continued as President of NBER and Geoffrey Carliner continued as Executive Director. Further information on NBER activities is available in the NBER *Reporter*, or from Geoffrey Carliner, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138.

DAVID KENDRICK, *Representative*

Report of the Representative to the Consortium of Social Science Associations

This note should stand as my report on the doings of the Consortium of Social Sciences on the Executive Committee of which I sit as AEA's Executive Committee member.

COSSA is continuing to lobby for additional funding for social science research. As always, it is impossible to measure effectiveness, because many separate influences are at work, but appropriations are up at NSF. On balance, I think COSSA continues to do a good job. In addition to direct lobbying for additional funding, COSSA is trying to provide services to members of Congress and their staffs, acting as a point of contact when they have enquiries and organizing small conferences around topics suggested to it by staff. People have been judiciously selected, and the conferences have gone well. In the long run, such efforts may be more important than direct lobbying.

COSSA has a new executive director, David Jenness, who replaced Roberta Miller, who in turn went to work at NSF in a position of some influence. He has done well in taking hold, and has received a laudatory letter from the COSSA Executive Committee on his work during his first year. COSSA finances appear sound. Additional universities and other groups continue to seek to affiliate with COSSA, and none, to the best of my knowledge, is dropping out. Given our profession's faith in the market, increasing membership, of course, is a good sign.

HENRY AARON, *Representative*

Report of the Committee on U.S.–China Exchanges

As mentioned in the Report of our Committee last year, 1984 could be considered as the year when modern economics became official in China. In 1985, the Chinese State Commission on Education (formerly the Ministry of Education) continued actively to strengthen economics education in China. Several important activities can be reported.

First, a Macroeconomics Workshop, sponsored by the State Commission on Education with partial financial support from the Ford Foundation, took place at the People's University in Peking from June 10 to July 20. Ninety-four persons from various universities and government research and planning organizations attended, with approximately 50 regular attendants and the remaining auditors. The instructors were John Taylor of Stanford, William Branson and Dwight Jaffee of Princeton, Richard Portes of the University of London, and Gregory Chow of Princeton, who also served as organizer and coordinator. The topics covered included macroeconomic theory and modeling by Taylor, open-economies macroeconomics by Branson, money and banking by Jaffee, macroeconomics of centrally planned economies by Portes, and applications to China by Chow. The quality of the students was on average better than those who attended the Microeconomics Workshop in 1984. The impact of the workshop is to enable the regular attendants coming from various universities to teach macroeconomics after their return to their respective schools. Also the attendants from government research and planning organizations will be able to use macroeconomics in their work. Represented were the Planning Commission, the combined Economic Research Centers of the State Council, the Economic Reform Committee of the State Council and the People's Bank.

Second, a year-round training program towards a master's degree in economics began in September 1985 at the People's University, also with financial support from the Ford Foundation. Forty-nine students enrolled in this program, having been selected mainly from seven major Chinese universities including Peking, People's, Nankai, Wuhan, Jilin, Fudan, and Xiamen. Daniel Suits from Michigan State and Kenneth Chan from McMaster taught micro and macro, respectively, in the fall semester of 1985. Leonid Hurwicz of Minnesota and Elizabeth Li of Temple will teach micro and macro, respectively, in the spring of 1986.

Sixty-two students from China, who had been recruited by the Ministry of Education in the fall of 1984, entered fifty-some graduate programs in economics in the United States and Canada in September 1985. These students are financially supported mainly by the universities accepting them, but some are supported by the Chinese government and the Ford Foundation. Again, in October 1985, examinations in mathematics and economics were given by the Chinese State Commission on Education to recruit students with the cooperation of a committee consisting of Edwin Mills, Sherwin Rosen, John Taylor, and Gregory Chow. These students are being recommended by the Committee to sixty-some U.S. and Canadian universities for admission in 1986.

It may be of interest to note that in July 1985, Premier Zhao Zhiyang asked me to invite foreign scholars to study problems of the Chinese economy. My first response was a suggestion to establish an economic data center at the People's University, a suggestion which the Premier promptly accepted. This center is now under preparation. I hope that in the future American scholars interested in studying the Chinese economy will take advantage of the material available in the center.

GREGORY C. CHOW, *Chair*

Report of the Committee on U.S.–Soviet Exchanges

Last year, for the first time since the current series of U.S.–Soviet exchanges began a decade ago, no economic symposium took place. This was due, primarily, to inevitable delays relating to the fact that Lloyd Reynolds, who had been the Chair since the beginning, stepped down, much to the regret of all those connected with the exchanges.

The ninth U.S.–Soviet Economic Symposium has just been scheduled for June 1986 to be held at Tufts University. The subject will be Aspects of the Economics of Agriculture. We feel that it is particularly appropriate to have a symposium on this subject at present because of the crucial problems facing agriculture in both the Soviet Union and the United States. An attempt will be made to publish those articles which meet refereeing standards.

FRANKLYN D. HOLZMAN, *Chair*

Report of the Committee on the Status of Women in the Economics Profession

One hundred years ago when the American Economic Association was founded, there were few women economists. Since that time there have been enormous changes in women's status in American society. Women are now permitted by law to vote, to attend the same schools as men, and to work in a variety of occupations outside the home. Currently, more than three-fifths of adult women (aged 20 to 64) are gainfully employed, compared to less than one-fifth a century ago. In spite of these changes, women still earn much less than men and lack the power and status traditionally associated with economic success. Their low earnings stem in large part from their concentration in low-paid occupations and their underrepresentation in most of the professions. In recognition of these facts, the American Economic Association established the Committee on the Status of Women in the Economics Profession (CSWEP) in 1972. Hopefully, by the time the American Economic Association celebrates its bicentennial in the year 2085, such a committee will no longer be needed.

This report summarizes changes in women's status within the profession over the past decade and describes the most recent activities of CSWEP. The overall trends on women's representation are generally positive for the decade as a whole and we would like to think that CSWP's activities contributed to some of that progress. It is also clear that women do not progress within the profession at the same rate that men do. We need to learn more about why this is so, and continue efforts to integrate them fully into the profession.

The Changing Status of Women Economists. As indicated in Figure 1, there has been a substantial increase in the number of women majoring in economics at the undergraduate level and in the number completing advanced degrees. Women now receive 34 percent of all BA degrees in economics, up from 22 percent a decade ago, and 18 percent of all PhD degrees, up from 11 percent over the same period. This has translated into considerable improvement at the entry level for the profession as well. In fact, women's share of all assistant professors has tended to mirror their share of PhD degrees over time. Progress to the top academic ranks of the profession, on the other hand, has been slow or nonexistent. Only 3 to 4 percent of all full professors were female in both 1974–75 and 1984–85.

It is interesting to compare the ten-year record to changes over the past four years (1981–85). Tables 1 and 2 contain data on a matched sample of institutions for these periods and, for comparative purposes, data from an unmatched sample (all that is available) for the preceding four years, 1977–81. One might hypothesize that there would have been a slowing of progress due to flagging interest in, and pressures for, affirmative action during this most recent period. However, most of the indicators presented in Tables 1 and 2 show continuing progress at rates comparable to earlier periods with two notable exceptions: the proportion of women BAs has levelled off and the proportion of women PhD students taking jobs in the academic sector appears to have dropped sharply. Since these are two critical points of entry into the profession, these data do not augur well for the future.

Closer inspection of the trends suggests that some recent progress is the result of 'pipeline effects' at work. For example, the sharp increase in women's representation at the MA degree level between 1981 and 1985 seems to mirror the sharp increase in their representation at the BA degree level between 1977 and 1981. Table 3 is an attempt to look at these pipeline effects more systematically by comparing women's representation at each level of the profession to a logically prior level four years earlier. If women progressed within the profession in the same way and at the same rate as men, the ratios in Table 3 would all eventually be close to 1.00. The fact that they are all well

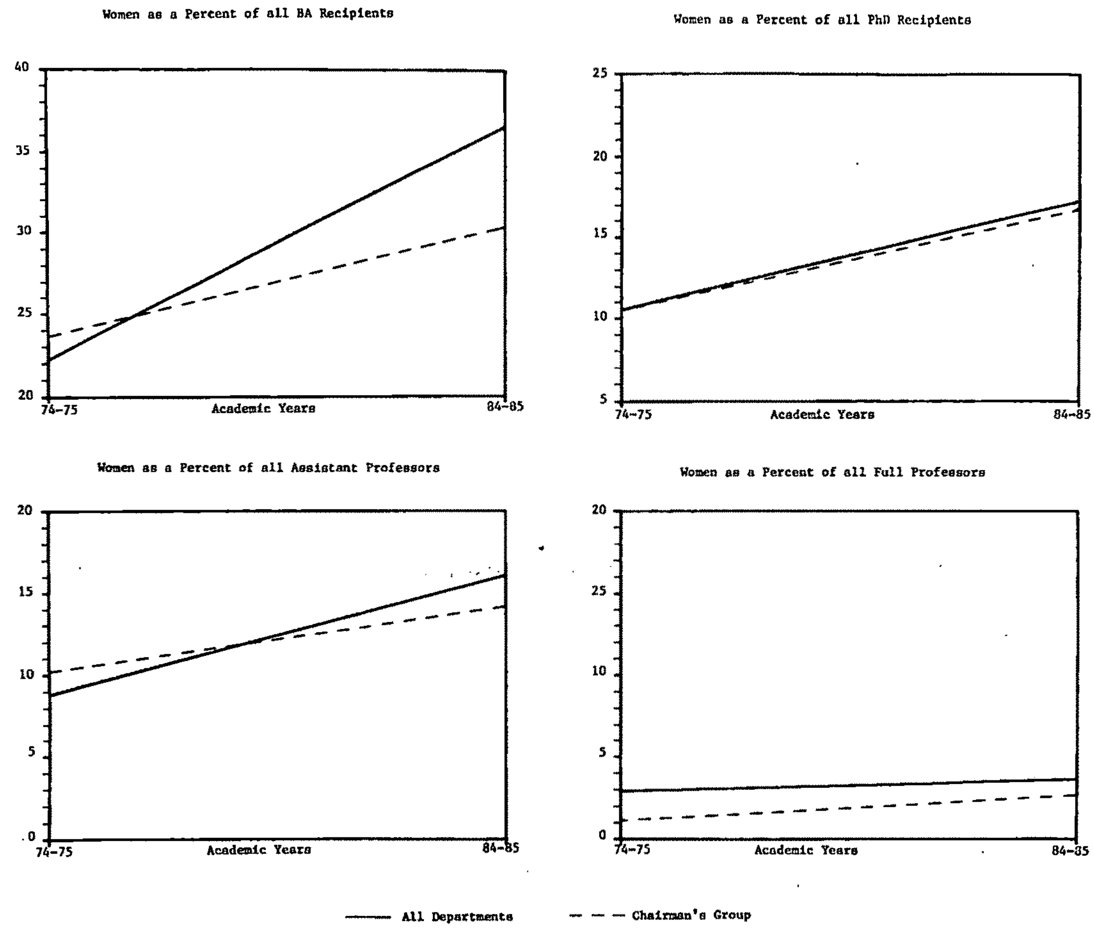


FIGURE 1. THE CHANGING STATUS OF WOMEN, 1975–85

below 1.00 indicates that the problem is not just a lack of women with the requisite prior training or experience. Moreover, the problem is more serious the higher one goes in the hierarchy. Women seem to succeed reasonably well (though not as well as men) in translating their educational credentials into a first job but much less well at moving up the ranks from assistant to associate to full professor.[1] (My impression is that they do better in nonacademic pursuits and this may be one reason the proportion of female PhD graduates entering the academic labor market has dropped.) The critical point appears to be promotion from assistant to associate professor suggesting that few women receive tenure. In this connection, it is interesting to note that the number of newly tenured people dropped between 1981 and 1985, but the proportion who were women went up, especially in the Chairman's Group (where no women received tenure in 1981). Another

[1]Since the data used to construct the ratios in Tables 3 and 4 are stocks (proportions at a point in time) rather than annual flows, it is possible for women to do as well as men in terms of hiring and promotion rates but still be poorly represented at the senior levels for many years since turnover is low in the higher ranks and women are the "new entrants."

TABLE 1—SELECTED DATA ON WOMEN'S STATUS IN THE ECONOMICS PROFESSION: ALL DEPARTMENTS

	1976–77		1980–81		1984–85	
	N = 343 Depts.		N = 200 Depts.		N = 200 Depts.	
	Total[a]	Percent Women	Total[a]	Percent women	Total[a]	Percent women
BA Degrees Awarded	10759	23.7	9975	35.0	11424	36.5
MA Degrees Awarded	1434	17.4	1149	22.7	1192	38.3
PhD Degrees Awarded	628	8.6	693	13.1	559	17.2
Employment[b]						
Asst. Professors	1294	8.5	889	11.9	876	16.1
Assoc. Professors	1017	4.5	698	4.4	825	7.9
Full Professors	1458	3.1	1314	3.0	1391	3.6
Changes in Academic Status[b,c]						
New Hires	339	9.1	231	10.4	201	21.4
Newly Tenured	131	9.2	75	8.0	71	9.9
Promoted to Rank	256	7.4	151	7.3	123	10.6
Graduate Students:[d]						
PhD Students	2389	14.3	4631	18.1	5090	20.4
MA Students	1080	17.3	2808	20.9	2061	26.3
Grad Students Receiving any Aid:[d]						
PhD Students	1802	14.4	3017	18.3	3320	21.1
MA Students	418	17.3	605	27.6	595	33.6
PhD Grads Employed as Economists:[e]	Male	Female	Male	Female	Male	Female
All Sectors	89.9	87.8	95.5	92.1	90.9	88.7
Educational Instit.	53.9	57.6	60.1	63.5	61.3	49.1

Source: Data for 1976–77 are from an unmatched sample of institutions responding to the Universal Academic Questionnaire. Data for 1980–81 and 1984–85 are from a matched sample.

[a]Total represents the sum of men and women and not necessarily the raw data totals.

[b]Includes both full-time and part-time professors.

[c]Only considers assistant, associate, and full professor slots.

[d]Includes both full-time and part-time students.

[e]Shown in percent. "All Sectors" includes: Educational Institutions, Business, Industry, Federal, State & Local Governments, Banking, Finance, Consulting, Research Institutions, Foreign Employment and International Agencies.

conclusion that can be drawn from Table 3 is that the *rate* of progress within the profession did not deteriorate between 1981 and 1985; indeed it appears to have improved somewhat at most levels.

In summary, we know that more and more women are acquiring the requisite training and experience to advance within the profession and that their ability to translate these into concrete advances within the academic community has probably improved somewhat. But women are still poorly represented, especially in the higher ranks. We do not know what factors lead to these gender differences. As always, one can advance both demand-side and supply-side reasons. One of CSWEP's priorities in the coming year will be to launch a more in-depth investigation of these factors, building on work done earlier in the committee's history by Barbara Reagan, Myra Strober, and others.

CSWEP Activities. CSWEP has traditionally maintained a roster of women economists. The data are usually updated annually and a hard copy mailed to all dues-paying members of CSWEP. Both the hard-copy version and on-line searches are available for use by employers and those interested in doing specialized research on women economists. There has been some debate within CSWEP about the utility of continuing the roster, particularly in light of the more frequent publication of the AEA Directory in recent years. But we have decided to continue publication of the roster for now since CSWEP has made a considerable investment in the basic data, because there are a very large number of women economists

TABLE 2—SELECTED DATA ON WOMEN'S STATUS IN THE ECONOMICS PROFESSION: CHAIRMAN'S GROUP

	1976–77		1980–81		1984–85	
	$N=43$ Depts.		$N=33$ Depts.		$N=33$ Depts.	
	Total[a]	Percent Women	Total[a]	Percent Women	Total[a]	Percent Women
BA Degrees Awarded	3196	21.2	3014	29.3	3562	30.4
MA Degrees Awarded	610	18.2	427	23.9	382	23.6
PhD Degrees Awarded	408	8.1	303	13.5	249	16.9
Employment[b]						
Asst. Professors	310	9.3	249	12.0	240	14.2
Asso. Professors	212	2.8	166	3.0	221	5.4
Full Professors	570	1.6	438	2.1	465	2.6
Changes in Academic Status[b,c]						
New Hires	88	3.4	50	10.0	55	12.7
Newly Tenured	12	8.3	23	0.0	13	15.4
Promoted to Rank	57	3.5	40	2.5	27	7.4
Graduate Students:[d]						
PhD Students	1951	14.5	2014	18.3	2240	20.2
MA Students	570	15.6	539	22.4	595	25.9
Grad Students Receiving any Aid:[d]						
PhD Students	1465	14.7	1329	18.5	1493	20.6
MA Students	189	15.3	183	26.8	180	33.9
PhD Grads employed as Economists:[e]	Male	Female	Male	Female	Male	Female
All Sectors	91.5	88.9	96.0	92.9	90.6	87.5
Educational Instit.	53.3	57.4	61.6	64.3	63.4	45.8

Source and fnn: See Table 1.

TABLE 3—UPWARD MOBILITY WITHIN THE PROFESSION: ALL DEPARTMENTS

	1980–81	1984–85
Women's Share of PhD and MA Degrees Awarded Relative to Share of BA Degrees Awarded Four Years Earlier	0.81	0.90
Women's Share of Assistant Professors Relative to Share of PhD and MA Degrees Awarded Four Years Earlier	0.80	0.84
Women's Share of Associate Professors Relative to Share of Assistant Professors Four Years Earlier	0.52	0.66
Women's Share of Full Professors Relative to Share of Associate Professors Four Years Earlier	0.67	0.82

Note: The above ratios are computed from the percentages in Table 1 and 2. Thus, if women earn 30 percent of all BAs, they might also be expected to earn 30 percent of (a smaller number of) all PhDs 4 or 5 years later if their rate of moving up the hierarchy were the same as men's. In this case, the ratio would be 1.00. Thus, this table attempts to measure, albeit crudely, whether women's rate of progress within the position has changed.

TABLE 4—UPWARD MOBILITY WITHIN THE PROFESSION: CHAIRMAN'S GROUP

	1980–81	1984–85
Women's Share of PhD and MA Degrees Awarded Relative to Share of BA Degrees Awarded Four Years Earlier	0.92	0.71
Women's Share of Assistant Professors Relative to Share of PhD and MA Degrees Awarded Four Years Earlier	0.85	0.72
Women's Share of Associate Professors Relative to Share of Assistant Professors Four Years Earlier	0.32	0.45
Women's Share of Full Professors Relative to Share of Associate Professors Four Years Earlier	0.75	0.87

Note: See Table 3.

who belong to CSWEP but not to the AEA, and because the CSWEP roster is a better tool for conducting targeted employment searches. In addition, updating the roster is a natural extension of the work entailed in maintaining a mailing list and sending out annual dues notices. We are extremely pleased that Joan Haworth has agreed to take on all of these tasks and we owe her and her staff a big debt of gratitude for all their hard work. A new roster is now being prepared and should be available in early 1986.

In addition to the roster, a major activity of CSWEP is publishing a newsletter three times a year. CSWEP spent considerable time this year discussing the purposes of the *Newsletter* (and implicitly, the purposes of CSWEP). A major issue is the extent to which the *Newsletter* should contain items of professional interest to women economists, whatever their field, and the extent to which it should feature material on gender-related research and the status of women generally. While we believe that both are important, the prevailing view of CSWEP was that more emphasis should be put on the former than the latter, and that any new editor should feel comfortable with this set of priorities. In this connection, I am very happy to report that Nancy Gordon, a new member of CSWEP, has agreed to take on the editorship of the *Newsletter*.

CSWEP is pleased to see an increasing number of women represented as office-holders and committee members of the AEA. For example, Elizabeth Bailey is a Vice President, Janet Norwood serves on the Executive Committee, Clair Brown was a member of last year's Nominating Committee, Marianne Ferber is on the Committee on Economic Education, Claudia Goldin and Susan Woodward are on the Editorial Board of the *American Economic Review* and Carolyn Shaw Bell on the Editorial Board of the *Journal of Economic Literature*.

We are particularly pleased that Alice Rivlin became President-elect in 1985. The President of the AEA serves as an *ex officio* member of CSWEP and CSWEP has generally tried to stay in contact with the President even though he or she does not normally attend our meetings. Rivlin demonstrated her particular interest in our efforts by accepting an invitation to attend CSWEP's first meeting this year. She saw three issues for possible CSWEP attention: (1) the process by which sessions and papers are chosen for the annual meetings; (2) the lack of upward mobility for women beyond the BA level in the profession and the possibility of doing some organized research on the reasons; and (3) using information from the Universal Academic Questionnaire (an outgrowth of earlier data collection efforts by this Committee) to learn more about the career patterns of both men and women within the profession.

The first issue was cogently addressed in an article by Cordelia Reimers in the CSWEP *Newsletter* (summer issue). In the article, she describes how the current process works and what women (or men) interested in getting on the program can do to improve their chances. CSWEP will continue to monitor the process, work with incoming Presidents to insure that women are represented on the program, and discuss possible modification of the procedures with the Executive Committee. We have written to this year's President-elect, Gary Becker, about our concerns.

The problem of upward mobility among women economists was highlighted in the first section of this report, and we are currently seeking foundation support for a small research project that would help us to learn more about why women have not made greater advances within the profession. Several foundations have expressed a willingness to consider support for the project and a number of good people have expressed interest in conducting the work. We would, of course, welcome any Executive Committee interest in extending such efforts to study the career patterns of economists more generally.

CSWEP has given attention to a number of other issues this year with various members of the board taking the lead responsibility. These activities have included consideration of a student prize in economics as a means of recognizing and encouraging young women to pursue further work in the field (Michelle White), compiling an on-line bibliography of women economists' publications (Mary Fish), arranging for a workshop on

the NSF economics grants program at the December meetings (Sharon Megdal), updating our information packet for those considering careers in economics (Sawhill), monitoring an ongoing project investigating gender bias in economics texts (Beneria), and investigating the extent to which women are appropriately represented on the editorial boards of various journals (Reimers). All of this is in addition to our usual activities of sponsoring sessions and get-togethers at the AEA and regional meetings. Particular thanks go to our regional chairs: Cordelia Reimers (CSWEP-Northeast), Sharon Megdal (CSWEP-West), Mary Fish (CSWEP-South), and Michelle White (CSWEP-Midwest).

Four CSWEP members' terms expire this year: Barbara Bergmann (past chair), Aleta Styers (past editor of the *Newsletter*), Cordelia Reimers, and Joseph Pechman. All have contributed substantially to CSWEP's work. They will be replaced by Beth Allen, Nancy Gordon, and Katharine Lyall.

ISABEL V. SAWHILL *Chair*

Report of the Committee on Economic Education

The Committee has worked closely with the Joint Council of Economic Education this past year on two developments. One is the issuance of a report by a committee of educators and economists recommending the kinds and amounts of training in economics and pedagogy needed for pre-college teachers to do a creditable job of teaching economics. The report, *Economic Education for Future Elementary and Secondary Teachers: Basic Recommendations*, is available from the *JCEE*, 2 Park Avenue, New York, NY 10016. The other is to set in motion plans for the development of an Advanced Placement Program in economics that would permit students who have access to outstanding high school courses in economics to qualify for advance college standing in economics as can now be done in many other subjects. A feasibility study by the College Board is now underway and should the results be favorable, the development of the economics test will commence sometime later this year. This means that the Program would be in place by the 1987–88 academic year.

The Committee is also interested in exploring the topic of graduate education in economics. The purpose would be to focus thinking on the nature and direction of graduate education, in the hope of stimulating research that would illuminate the effectiveness of the various dimensions of graduate education. The Committee prepared and submitted a proposal to the Executive Committee for funding a Symposium of Graduate Education with the papers to be published in the *Journal of Economic Education*; the Executive Committee decided not to fund the proposal. The Committee is now deciding what course of action to pursue.

An evaluation of the Teacher Training Program was completed during the past year. A survey of former participants indicated that summer workshops were regarded as highly valuable in stimulating greater interest in teaching and providing assistance for improving teaching skills. The Resource Manual was also judged to be valuable but in need of revision; a variety of useful suggestions for revision emerged. Finally, it was generally recognized that the video tapes are dated and no longer effective; the universal recommendation was that they be redone professionally. Committee members are now preparing a proposal to obtain funding to revise the Resource Manual and to produce a new set of video tapes. Consideration is also being given to how best promote the use of these materials and the improvement of teaching, particularly among new entrants into the teaching ranks.

W. LEE HANSEN, *Chair*

New

Strategic Trade Policy and the New International Economics

edited by Paul Krugman

These original essays bring the practical world of trade policy and of government and business strategy together with the world of academic trade theory. They focus in particular on the impact of changes in the international trade environment and on how new developments and theory can guide our trade policy.
$12.50 original in paperback (cloth $27.50)

Market Structure and Foreign Trade

Increasing Returns, Imperfect Competition, and the International Economy

Elhanan Helpman and Paul R. Krugman

"A landmark book that we will each need in our library."
–Paul Samuelson
$22.50

New

Superfairness

Applications and Theory

William J. Baumol

Superfairness is a stunning display of applied and theoretical microeconomics, addressing the problem of how to analyze the fairness of the distribution of resources, product, income, and wealth that arises from economic decisions.
$20.00

New

Microtheory

Applications and Origins

William J. Baumol

These essays provide an engaging intellectual history of one of the leading figures in the field of economics.
$35.00

New

Dollars, Debts, and Deficits

Rudiger Dornbusch

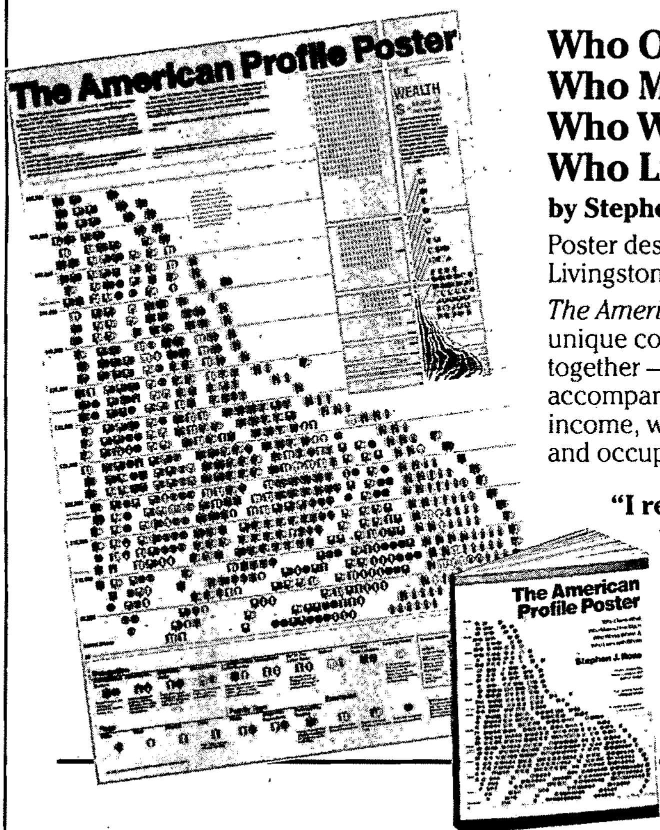
These essays address most if not all of the key current policy issues in open economy macroeconomics: the strong dollar, LDC debt problems, and deficit financing. Dornbusch brings a common political economy perspective to bear on the issues, revealing that more than ever, modern macroeconomics is useful as a framework for active policy.
$20.00

Inflation, Debt, and Indexation

edited by Rudiger Dornbusch and Mario Henrique Simonsen

"Essential reading for those concerned with indexation and economic stabilization issues . . . this single volume provides the 'state of the art' of a much debated issue."–*Finance & Development*
$9.95 paper

The American Profile Poster



Who Owns What, Who Makes How Much, Who Works Where, and Who Lives With Whom

by Stephen J. Rose

Poster designed by Dennis Livingston and Kathryn Shagas

The American Profile Poster is a unique color wall chart that brings together—along with its remarkable accompanying book—data on the income, wealth, race, marital status and occupation of all Americans.

"I recommend it wholeheartedly."
—Robert L. Heilbroner

Overall size 21″ x 32″
$8.95, available now.

THE BIG BOYS

Styles of Corporate Power

by Ralph Nader and William Taylor

The book *The Wall Street Journal* said "promises to be a doozy." Incisive profiles of nine powerful business leaders (David Roderick, Roger Smith, Felix Rohatyn, others) reveal how and *why* big business *is* the life of America today. Available in June.

THE MONEY MANDARINS

The Making of a New Supranational Economic Order

by Howard M. Wachtel

"A brilliantly written, absorbing, alarming and entirely true account of a financial condition that embraces the entire Western world."
—Robert L. Heilbroner

"A fascinating story with important new insights." —Richard J. Barnet

Available in June.

To order, call toll-free (credit cards only)
1-800-638-6460

PANTHEON BOOKS
201 East 50th St., N.Y. 10022

CLAIRE BRETÉCHER. © LE NOUVEL OBSERVATEUR

WRITING FOR SOCIAL SCIENTISTS

How to Start and Finish Your Thesis, Book, or Article

Howard S. Becker

With a Chapter by

Pamela Richards

Becker offers eminently useful suggestions for ways to make social scientists better and more productive writers. Throughout, Becker's focus is on the elusive work habits that contribute to good writing, not the more easily learned rules of grammar or punctuation.

Paper $6.95   180 pages
Library cloth edition $20.00
Chicago Guides to Writing, Editing, and Publishing

STRIKING A BALANCE

Making National Economic Policy

Albert Rees

"[Rees] examines our institutions of economic policy . . . [and] explains how each deals with the other, how they all nurture and defy the forces at work in the economy; how they juggle the conflicting goals of equality and efficiency, full employment and stable prices. . . . There's been no better recent primer on economic policy."—Peter Kilborn, *New York Times Book Review*

Paper $5.50   128 pages

CHICAGO

UNION RELATIVE WAGE EFFECTS

A Survey

H. Gregg Lewis

"Lewis surveys and integrates an incredibly large number of separate studies, corrects various errors in computing and in the coding of data, and also corrects many errors in analysis and in interpretation. The result is the most important analysis available on the impact of unions on relative wages in the United States."—Gary Becker, University of Chicago

Cloth $37.50   248 pages   42 tables

INFLATION, EXCHANGE, RATES, AND THE WORLD ECONOMY

Lectures on International Monetary Economics

Third Edition

W. M. Corden

Updated and substantially revised, with new chapters on the international transmission of economic disturbances, the international macrosystem, and macroeconomic policy coordination, Corden's book remains one of the most useful, comprehensive accounts of international monetary economics.

Paper $8.95   204 pages
Library cloth edition $22.50

The University of CHICAGO Press

5801 South Ellis Avenue, Chicago, IL 60637

MATHEMATICAL AND STATISTICAL PROGRAMMING PACKAGE FOR YOUR IBM PC

FAST • EASY TO USE • POWERFUL

GAUSS™

YOU'VE NEVER SEEN ANYTHING LIKE IT!

GAUSS is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines **speed, power, and ease of use** in one amazing program.

GAUSS allows you to do essentially **anything you can do with a mainframe** statistical package — and a lot more.

Personal computers are **friendly, convenient, and inexpensive.** So is **GAUSS**. **GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS**, you may never use a mainframe again.

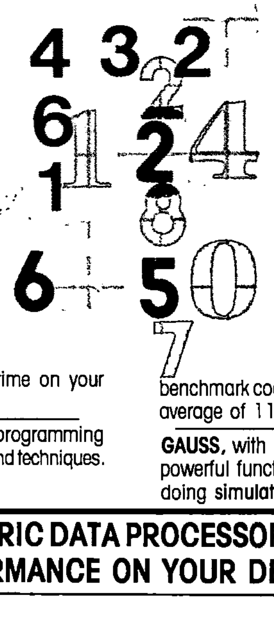
GAUSS comes with programs written in its matrix programming language that allow you to do most **econometric procedures**, including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is **no limit** on the number of observations.

GAUSS will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the **maximum likelihood** estimates of a binary logit model with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

GAUSS allows you to do **complicated statistical procedures** that you would never imagine trying on a mainframe. It is easy **to program** almost any routine, and **GAUSS** is so **fast** that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

GAUSS is an **excellent teaching tool.** It makes programming easy and allows students to focus on concepts and techniques.



If you can write it mathematically, you can write it in **GAUSS.** Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

GAUSS is **10-15 times faster** than other programs that use the 8087, and **15-100 times faster** than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional **powerful numerical operators and functions** — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

GAUSS has **state-of-the-art numerical routines and random number generators.**

GAUSS is **extremely accurate.** It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coeficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

GAUSS, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing **simulations.**

GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.

SPECIAL INTRODUCTORY OFFER
With 30 Day Money Back Guarantee .... Reg. 395.00 **$250.00**

GAUSS requires an IBM PC with at least 256K RAM, an 8087 NDP, 1 DS/DD disk drive, DOS 2.0 (or above).

IBM is trademark of IBM Corporation

Call or Write **APPLIED TECHNICAL SYSTEMS**
P.O. Box 6487, Kent, WA 98064
(206) 631-6679